

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/109386>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

Recurrence Relationships and model monitoring for Dynamic Linear Models

Parmaseeven Pillay Veerapen

Thesis submitted for the degree of
Doctor of Philosophy

Department of Statistics

University of Warwick

December 1991



TO

(a) posthumously, my Father and my Mother who passed away whilst I
was away;

(b) Linda and Maeva who had to bear with me for so long.

<u>CONTENTS</u>	<u>Page No.</u>
<u>NOTE TO THE READER</u>	(i)
<u>ACKNOWLEDGEMENTS</u>	(ii)
<u>SUMMARY</u>	(iii)
1. <u>INTRODUCTION</u>	1
2. <u>THE THEORETICAL FRAMEWORK - A BRIEF REVIEW</u>	5
2.1 The Bayesian Approach to Statistics	5
2.1.1 Introduction	5
2.1.2 Probability	5
2.1.3 Inference	7
2.1.4 Decision Theory/Analysis	9
2.2 Bayesian Forecasting and Dynamic Models	11
2.2.1 Dynamic Models	11
2.2.2 Bayesian Forecasting	13
2.2.3 Comments	14
2.3 The Dynamic Linear Model	14
2.3.1 Model Definition	14
2.3.2 Updating Equations, Observational Variance V_t Known	16
2.3.3 The Case of an Unknown Constant Observational Variance $V = \sigma^2$	18
2.3.3.1 Model Definition	18
2.3.3.2 Updating Equations	19
2.3.4 The DLM and Conditional Independence	21
2.4 Model Design and Specification	22
2.4.1 Introduction	22
2.4.2 Observability	23
2.4.3 Superposition	24
2.4.4 Discounting	25
2.5 Dealing with Model Uncertainty	27
2.5.1 Introduction	27
2.5.2 Filtering and Smoothing	28
2.5.3 Intervention	28
2.5.4 Monitoring	31
2.5.5 Multi-Process Models	32
2.6 Extensions to the NDLM	34
2.7 Final Comments on the DLMs	36
2.8 Bayesian Forecasting and the Classical Approach	36

3.	<u>GENERAL RESULTS FOR NORMAL DYNAMIC MODELS</u>	38
3.1	Introduction	38
3.2	Standard Normal Results for Incorporating Information	39
3.3	Leverage of an Observation on the States	41
3.4	Deletion of Information	43
	3.4.1 The Case of Known Variance	43
	3.4.2 The Case of Unknown Variance	48
3.5	The Special Case of the Dynamic Linear Model	50
4.	<u>CONDITIONAL INDEPENDENCE FOR NORMAL DISTRIBUTIONS</u>	52
4.1	Introduction	52
4.2	Important Results	52
4.3	Comments and Diagrammatic Illustration	57
	4.3.1 Conditional Independence Diagrams	57
	4.3.2 Incorporating Information	59
4.4	Routine Learning and Updating in the NDLM	61
4.5	Retrospective Analysis in the NDLM	63
5.	<u>DELETION OF INFORMATION</u>	67
5.1	Introduction	67
5.2	Deletion of a Single Observation	67
	5.2.1 Preliminary Comments and Notation	67
	5.2.2 Main Results	69
5.3	Finite Truncation Model	72
5.4	Deletion of a Set of k Observations	73
	5.4.1 Introduction	73
	5.4.2 Methodology/Procedure	73
	5.4.3 The Case of Unknown Constant Variance	75
5.5	The Case of Stochastic Variance	76
6.	<u>THE SPECIAL CASE OF DISCOUNT WEIGHTED REGRESSION AND APPLICATIONS</u>	81
6.1	Introduction	81
6.2	Discount Weighted Regression	82
	6.2.1 Definition	82
	6.2.2 Equivalence of DWR with the DLM	84
	6.2.2.1 Definition and Notation	83
	6.2.2.2 Theorem 6.1	86
	6.2.2.3 Conclusion	90
6.3	Incorporation of Information	90

6.4	Deletion of Information	93
6.4.1	Deletion of One Observation y_i	93
6.4.2	Deletion of a Set of Observations	94
6.5	Applications	95
7.	<u>CUSUMS AND MODEL MONITORING - A REVIEW</u>	106
7.1	Introduction	106
7.2	General Background	106
7.3	The Cusum	109
7.4	The V-Mask	113
7.5	Other Aspects of Cusums	115
7.6	Cusum and Model Monitoring	117
8.	<u>SEQUENCES OF SPRT's (SSPRT's)</u>	119
8.1	Introduction	119
8.2	The Sequential Probability Ratio Test (SPRT)	119
8.3	Sequences of SPRT's (SSPRT)	122
8.3.1	Definition	122
8.3.2	The Case of the Normal Mean (With Known Variance)	124
8.4	The Exponential Family - A General Result	126
8.4.1	Definition of the Exponential Family	126
8.4.2	The General Result	127
8.5	Special Cases	129
8.5.1	The Binomial Model	129
8.5.2	The Poisson Case	131
8.5.3	The Normal Variance Case	132
8.5.4	The Case of t-distribution	135
8.6	Comments	138
9.	<u>A BAYESIAN DECISION APPROACH TO CUSUMS</u>	139
9.1	Introduction	139
9.2	The Loss Function	139
9.3	The Normal Mean Case	143
9.3.1	The Model	143
9.3.2	Learning on μ	143
9.3.3	The Bayesian Decision Scheme $BD(a,b)$	144
9.4	Bayesian Decision Scheme : A General Result	148
9.4.1	Introduction	148
9.4.2	General Result	148
9.4.3	Comments	152

10.	<u>SPECIAL CASES AND APPLICATIONS</u>	154
10.1	Introduction	154
10.2	The Binomial Case	154
10.3	The Poisson Case	158
10.4	The Gamma Case and the Normal Variance	159
	10.4.1 The Gamma Model	159
	10.4.2 Monitoring of the Normal Variance	162
10.5	Monitoring a Normal Mean with Unknown Variance	163
10.6	Applications : The Case of Monitoring a Normal Variance	166
11.	<u>CONCLUSION</u>	170
	<u>REFERENCES</u>	174

NOTE TO THE READER

Observation vectors are considered throughout in this thesis. However, in two or three cases, for clarity in the presentation of the ideas, the case of univariate observations is considered initially (for example, Chapter 2, Sections 2.3.1, 2.3.2, 2.3.3; 2.4.2; Chapter 6, Section 6.2.1). In these cases, the vectors (usually the parameters/states, system variances etc.) are underlined to distinguish them from the scalars (usually the observation and the observational variance).

(ii)

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor P J Harrison for his expert guidance, for the stimulating discussions I have had with him and for his sustained encouragement.

I thank also D V Lindley for having raised the question concerning the relationship between Cusums and Bayesian decisions.

My thanks go also to my fellow research students for their friendship and support, with particular mention to Klaus Vasconcellos for his help in computing.

I thank Mandy Broom for having typed the thesis, and for having made sure that I meet the required deadline.

Finally I would like to thank the University of Mauritius for having agreed to provide me with the opportunity to carry out this research and the Association of Commonwealth Universities for the financial support.

(iii)

SUMMARY

This thesis considers the incorporation and deletion of information in Dynamic Linear Models together with the detection of model changes and unusual values. General results are derived for the Normal Dynamic Linear Model which naturally also relate to second order modelling such as occurs with the Kalman Filter, linear least squares and linear Bayes estimation.

The incorporation of new information, the assessment of its influence and the deletion of old or suspect information are important features of all sequential models. Many dynamic sequential models exhibit conditional independence properties. Important results concerning conditional independence in normal models are established which provide the framework and the tools necessary to develop neat procedures and to obtain appropriate recurrence relationships for data incorporation and deletion. These are demonstrated in the context of dynamic linear models, with particularly simple procedures for discount regression models.

Appropriate model and forecast monitoring mechanisms are required to detect model changes and unusual values. Cumulative Sum (Cusum) techniques widely used in quality control and in model and forecast monitoring have been the source of inspiration in this context. Bearing in mind that a single sided cusum may be regarded essentially as a sequence of sequential tests, such a cusum is, in many cases, equivalent to a Sequence of Sequential Probability Ratio Tests in many cases, as for example in the case of the Exponential Family. A relationship between cusums and Bayesian decision is established for a useful class of linear loss functions. It is found to apply to the Normal and other important practical cases. For V-mask cusum graphs, a particularly interesting result which emerges is the interpretation of the distance of the V vertex from the latest plotted point as the prior precision in terms of a number of equivalent observations.

This thesis deals with certain aspects of sequential dynamic modelling in a Bayesian framework. In particular amongst other things, it considers the tracking of unusual values and outliers or suspect data. In many cases, these are manifestations of model uncertainty. Appropriate model monitoring mechanisms are necessary to deal with such cases and are presented in this thesis. Further, the identification or detection of suspect data may lead to a decision to delete them, especially in cases where they may influence unduly current forecasts. The thesis deals in some detail with the deletion of one observation and a set of observations from a given data set. Corresponding recurrence relationships for the first two moments of the joint distribution of the states are developed which dual those developed for the incorporation of information.

First, the theoretical framework within which the work in this thesis has been carried out is presented in Chapter 2: the Bayesian approach to probability, inference and decision theory/analysis including the subjectivistic view of probability. Particular emphasis is placed on dynamic modelling and Bayesian forecasting, and, in particular, on the Dynamic Linear Model developed by Harrison and Stevens (1976).

Thereafter, the thesis is divided into two parts: the first part, Chapters 3 to 6, covers the incorporation and deletion of information in dynamic models in some detail, and the second part, Chapters 7 to 10, deals with the development of appropriate model monitoring mechanisms making use of, amongst others, the cumulative

sum technique as a source of inspiration.

In sequential dynamic modelling, new information is combined with historic information both in order to update forecasts, and to reassess what happened in the past. Chapter 3 presents general results for normal dynamic models on the incorporation and deletion of information in a fairly general setting, when the observational variance is known only up to a scalar factor. Strong conditions like conditional independence properties are not required for these results. Equally in Chapter 3, an expression for the leverage of an observation on the states of a normal model is also derived.

A key property of normal dynamic models is conditional independence. Powerful results generally applicable to normal models are developed in Chapter 4 which includes also the discussion of the application to dynamic linear models. Conditional independence diagrams which are special cases of undirected graphs are presented both for the general case and the Dynamic Linear Model (DLM). Using these results, the incorporation of information in normal dynamic models is illustrated in two particular cases. The routine updating and learning procedures are obtained in a very straightforward application of the results, and, the retrospective analysis is seen as an example of incorporation of information whereby the full historic state distribution is obtained given up-to-date information. Its calculation, and in particular that of its first two moments, are seen as an immediate consequence of conditional independence.

The deletion of information is considered in Chapter 5,

sum technique as a source of inspiration.

In sequential dynamic modelling, new information is combined with historic information both in order to update forecasts, and to reassess what happened in the past. Chapter 3 presents general results for normal dynamic models on the incorporation and deletion of information in a fairly general setting, when the observational variance is known only up to a scalar factor. Strong conditions like conditional independence properties are not required for these results. Equally in Chapter 3, an expression for the leverage of an observation on the states of a normal model is also derived.

A key property of normal dynamic models is conditional independence. Powerful results generally applicable to normal models are developed in Chapter 4 which includes also the discussion of the application to dynamic linear models. Conditional independence diagrams which are special cases of undirected graphs are presented both for the general case and the Dynamic Linear Model (DLM). Using these results, the incorporation of information in normal dynamic models is illustrated in two particular cases. The routine updating and learning procedures are obtained in a very straightforward application of the results, and, the retrospective analysis is seen as an example of incorporation of information whereby the full historic state distribution is obtained given up-to-date information. Its calculation, and in particular that of its first two moments, are seen as an immediate consequence of conditional independence.

The deletion of information is considered in Chapter 5,

where in Section 5.2 dual recurrence relationships are derived for the elimination of a single observation. This immediately leads to a simple operation for finite truncated models where the forecasts are based on at most the last k time periods. Then Section 5.4 generalises to the deletion of any set of past observations with a very elegant procedure for revising distributions. The resulting jackknifed posterior state and various predictive distributions provide the basis for future derivation of relevant diagnostics such as those advocated by Smith and Pettit (1985), Bernardo (1985) and Johnson and Geisser (1983). Chapter 5 ends by considering a stochastic variance model together with a method of deriving its current jackknifed distribution.

The subset of discount weighted regression dynamic models is defined in Chapter 6 since they exhibit very neat procedures and provide a link with static models and least squares procedures. The equivalence between discount weighted regression (DWR) and the normal DLM is first established; thereafter simple neat results corresponding to those for normal dynamic models obtained in Chapters 4 and 5 are derived for DWR and Exponentially Weighted Regression (EWR), itself a special case of DWR.

The second part of the thesis dealing with model monitoring starts with Chapter 7. The cusum technique, as used in industrial quality control, has been to a great extent a source of inspiration for model monitoring in general and in the present work. Thus a review of the ideas related to the cusum technique is presented in Chapter 7, as a useful background to the ideas to be developed in the subsequent chapters. In particular, the two major ways in which

cusums operate are defined: the Cusum Decision Scheme and the V-mask Cusum Graph.

Moreover, a single-sided cusum may be regarded essentially as a sequence of sequential tests. Chapter 8 develops results for a monitoring mechanism using a sequence of sequential probability ratio tests (SSPRTs) whereby the Bayes' Factor is made use of. It is then established that the Cusum Decision Scheme is, in fact, a special case of SSPRTs, particularly for exponential family models.

In Chapter 9, a link between Cusums and Bayesian Decision Theory is established. A loss function, linear in the parameter being monitored, is constructed relative to the loss of continuing the sample run. After looking at the Normal mean case, a more general result is obtained which directly relates to many of the application cases. The loss function corresponds to the defining characteristics of a Cusum Decision Scheme and in particular, for V-mask Cusums, the distance of the vertex of the mask from the latest plotting point is the prior parameter precision expressed in terms of equivalent observations.

The special cases and applications of the results developed in Chapter 9 are then considered in some detail in Chapter 10. The cases of Binomial, Poisson and Gamma distributions are discussed where, for Gamma distributions, the case of Normal variance monitoring is included. The application of these ideas to a data set is then presented.

Chapter 11 concludes the thesis with a brief discussion of research perspectives emerging out of the ideas and results developed, discussed and presented throughout the thesis.

CHAPTER TWO

THE THEORETICAL FRAMEWORK - A BRIEF REVIEW

2.1 The Bayesian Approach to Statistics

2.1.1 Introduction

Ever since the Reverend T. Bayes developed in 1763 the famous theorem which bears his name, Bayesian statistics have evolved along many strands. The view of probability and statistical inference as expressed in Bayes' theorem (Biometrika, 45, 1958) and as thereafter developed is fundamentally different to the various schools of thought (including the dominant classical/frequentist view). It is perhaps not surprising that conflicting lines of thought do eventually emerge. In this Chapter, the Bayesian approach incorporating the subjectivistic view of probability is briefly reviewed; this is the school of thought adopted in this thesis.

Any global approach to statistics, like the Bayesian one, would consist of three main components necessary to tackle any statistical problem (or even other problems as well): probability, inference and decision theory/analysis. The Bayesian view of each of these three components is now briefly considered.

2.1.2 Probability

As De Finetti (Theory of Probability, 1974), who comprehensively developed the subjectivistic view of probability,

put it, 'Probability does not exist', i.e. probability does not exist objectively outside a person. He added, 'The only relevant thing is uncertainty - the extent of our knowledge and ignorance'. Probability is a description of one's uncertainty about the world given one's state of knowledge and ignorance. It is one's degree of belief in a certain proposition about the world or some aspect of the world, again subject to one's knowledge and ignorance.

That degree of belief would then be modified as one's state of knowledge is modified, for example through information from new sources or through some form of data collection or through a better understanding of whatever is being described or studied. This leads invariably to the assertion that probability can only be conditional and hence to the principle of conditionality inherent in the Bayesian view of probability. As H. Jeffreys (*Theory of Probability*, 1961), though not a subjectivist, aptly puts it, 'Our fundamental idea will not be simply the probability of a proposition p , but the probability of p on data q .'

It is very pertinent to note that Kolmogoroff (1937) developed probability theory axiomatically and that Bayes' Theorem accords with these axioms, so that the Bayesian approach is essentially simply that of probability.

That view of probability underpins the Bayesian approach to inference in general and to modelling in particular. It helps to develop a unified view of statistics as a whole.

2.1.3 Inference

Bayesian inference is essentially the study of how degrees of belief in a given proposition are altered by data, by making use of Bayes' Theorem.

When studying a process or phenomenon or any aspect of the world, one develops models to describe reality. From prevailing (including historical) information or knowledge, prior to any data collection, one formulates a degree of belief in a proposition concerning the parameters of the given model. Then data are collected and the information obtained from the data is expressed through the likelihood function. There is a need now to evaluate and update the degree of belief in the light of new information. Bayes' Theorem precisely provides the mathematical formulation for combining previous knowledge as contained in the prior distribution with new knowledge as contained in the likelihood to give the posterior degree of belief as expressed by the posterior distribution. Hence, the crucial role of Bayes' Theorem which formally states:

Posterior distribution = Prior distribution \times Likelihood function, i.e.

$$p(\theta | y) = p(\theta) l(\theta | y)$$

where θ : parameter vector.
 y : set of observations,
 $l(\cdot | \cdot)$: the likelihood function.

It clearly follows that Bayes Theorem exhibits a sequential characteristic. Let y_1 be a set of observations at time $t = t_1$. Then from Bayes' Theorem,

$$p(\theta | y_1) \propto p(\theta) l(\theta | y_1)$$

And let y_2 be a second set of observations independently taken at time $t = t_2$, $t_1 < t_2$. Then Bayes' Theorem would give

$$\begin{aligned} p(\theta | y_1, y_2) &= p(\theta) l(\theta | y_1, y_2) \\ &\propto p(\theta) l(\theta | y_1) l(\theta | y_2) \\ &\propto p(\theta | y_1) l(\theta | y_2) . \end{aligned}$$

Thus, $p(\theta | y_1)$, the posterior relative to data set y_1 , becomes the prior at time t_2 . This sequential characteristic of Bayes' Theorem reveals the essence of Bayesian inference; it is all about continuous learning from experience, and thus continuously improving one's degree of belief.

The prior, the likelihood function and the posterior are the central ideas of Bayesian inference. It is not possible to discuss the details of any of these three inputs here. But it is, however, important to mention Bayes' Postulate and the likelihood principle which is another key idea of Bayesian inference. In very simple terms, the likelihood principle states that all information about a set of hypotheses from data and sampling models alone, is

expressed through the likelihood function.

Bayes' postulate deals with the prior; in particular it is concerned with how ignorance is formulated a priori, using a uniform distribution. There are many objections which have been formulated regarding this postulate. Many of these are not concerned so much with the ideas themselves but with their application and, in particular, at the possibility of their abuse. Monitoring model performance is the key to detecting such abuse or misconceived models, and the thesis is particularly concerned with this topic, as developed in Chapters 7 to 10.

Finally, especially in the context of the thesis, it is to be noted that using the posterior distribution, predictive distribution can be easily obtained so that predictions follow in a fairly straightforward fashion, expressed by probability distributions.

2.1.4 Decision Theory/Analysis

Over the last thirty years, major developments in decision theory/analysis from a Bayesian perspective have been achieved with major contributions from Raiffa and Schlaifer (Applied Statistical Decision Theory, 1961), Raiffa (Decision Analysis, 1968), De Groot (Optimal Statistical Decisions, 1970), Lindley (Making Decisions, 1985), Berger (Statistical Decision Theory and Bayesian Analysis, 1985), J.Q. Smith (Decision Analysis - A Bayesian Approach, 1988).

The Bayesian approach to decision theory and analysis is

conceptually very straightforward. The key idea is the loss function, $L(\theta, d)$, which expresses mathematically the loss incurred when a decision d is taken and θ turns out to be the outcome. In business problems, this loss will be measured usually in monetary terms; in other cases, the units of measurement may be different. There are various types of loss function and there exist characteristics which loss functions should possess to make them efficient.

Usually, the best way to construct loss functions which are reasonable is to carry out a utility analysis. In practice, though, there exist some standard loss functions which are quite useful. For example, the quadratic or squared-error loss function defined by

$$L(\theta, d) = (\theta - d)^2$$

or some linear loss function as defined by the absolute loss function. These are examples of unbounded loss functions. An example of a bounded loss function is the step loss function,

$$L(\theta, d) = \begin{cases} 0 & , |\theta - d| \leq b \\ 1 & , \text{otherwise} \end{cases}$$

Using Bayesian inference, the posterior distribution of θ , $p(\theta|y)$, can be obtained and, in turn, used to obtain the posterior expected loss. Then a Bayes rule (or decision) is that rule which minimises the posterior expected loss. Such a rule may not be necessarily unique. The relevance of boundedness for loss

functions, now becomes evident. In the case of unbounded loss functions the expected loss may be infinite in which case a Bayes decision cannot be obtained.

Whilst the above ideas provide the framework to solve business problems and similar types of problems in a logical manner, Bayesian decision theory does overlap with Bayesian inference. In fact, some people do argue that inference problems do not really exist; for example, the choice of an inference can be viewed as a decision problem. This is beyond the scope of the present discussion; moreover, an example of the overlap mentioned above is obviously the problem of hypothesis testing. For example, in this thesis, using Bayesian decision approach, a different view of sequential procedures and hypothesis testing is developed in the context of model monitoring.

2.2 Bayesian Forecasting and Dynamic Models

2.2.1 Dynamic Models

Any sensible model is designed with the objective of representing reality or a system or simply a process as faithfully as possible, and of estimating or predicting when necessary, and, almost always, to ultimately help in taking correct decisions and actions with regard to that reality or system or process under investigation. There has long been the assertion that a single model is the true one to achieve those objectives. However, the Bayesian approach to modelling does not lend itself to such simplistic perception of modelling. According to that approach and in line with the Bayesian

approach to probability and inference, a model is a way of viewing reality/system/process and its context.

It is expected, therefore, that at a particular time there will almost certainly be different competing views of the system/process and its context. Most often, if not always, there will be a 'dominant' view which prevails. In fact, in many cases, that 'dominant' model will prevail over a fairly long period of time; it is then referred to as a routine way of viewing the system/process. There is thus a source of model uncertainty which demands the definition of a set of models as opposed to one 'true' model which represents a given system/process. It may well be that those modellers, believing in one 'true' model, tend to identify, in spite of themselves, the 'dominant' model as the eternal one, which, of course, leads invariably to contradiction and model failures.

Another important source of model uncertainty is of crucial importance: it is based on our understanding of reality. The world with its myriad of processes and systems is dynamic: they are in a state of perpetual motion and change, usually somewhat slow so as to be not easily perceptible and sometimes sudden and pronounced. A model has to reflect that state of affairs, a source of uncertainty to the modeller due to the passage of time. Thus because the model form is only locally appropriate in time, the parameters θ are defined so as to reflect that state of perpetual slow change. The indexing of the θ 's by t , i.e. θ_t and the modelling of their 'behaviour' by a single random walk sum up these characteristics.

A final major source of model uncertainty is due to the

fact that change does not occur slowly all the time, as mentioned briefly in the previous paragraph. At a time in the future, there may be sudden and sharp changes in a process or system so that the corresponding model needs to be changed drastically. Then the whole model form changes and may even involve different input variables.

These major sources of model uncertainty need to be built in the model design and definition. Dynamic models do precisely that. In simple practical terms, they may be generally defined as 'sequences of sets of models', as expressed by M. West and P.J. Harrison (1989a).

2.2.2 Bayesian Forecasting

Bayesian Forecasting (Harrison and Stevens, 1976) uses Bayesian inference to study processes and systems through dynamic models. It develops appropriate learning procedures for parameters, variance and model definition, so that we can learn from the past, combine the past with existing data and then predict the future.

Bayesian forecasting therefore helps, through a whole set of tools and techniques, to develop an efficient learning system which is responsive to the ever changing reality. That is why we refer to the development of a forecasting system as opposed to a forecasting model, for the forecasting system caters for interaction between the modeller/forecaster, the model, and the process under investigation together with the environment in which the process operates.

In line with the Bayesian view of probability, the forecasts for parameters and observations are given in terms of probability distributions, from which moments and other characteristics of interest can be derived.

2.2.3 Comments

In what follows in this chapter (except section 2.8), the implementation of the ideas (expressed above and in section 2.1) in terms of model definition and design, learning procedures, model monitoring mechanism etc. is reviewed. The ideas developed so far thus represent a sort of background canvas on which the artist (i.e. the modeller) has had his work done.

It is to be noted that although the theoretical framework has the potential for a wide range of applications, we are more concerned here with time series analysis and forecasting.

2.3 The Dynamic Linear Model

2.3.1 Model Definition

The Dynamic Linear Model, DLM, characterised by the quadruple $\{F_t, G_t, V_t, W_t\}$, encapsulates impressively the ideas on modelling briefly discussed in section 2.3. We shall deal mainly with the Normal Dynamic Linear Model, NDLM, because, on the one hand, it makes the presentation simple and, on the other hand, the thesis, especially the 'early' part, deals mainly with NDLMs. Thereafter, further developments of and extensions to NDLM will be highlighted.

In the quadruple referred to above, F_t and G_t are respectively known $(n \times r)$ and $(n \times n)$ matrices; and V_t , W_t are respectively $(r \times r)$ and $(n \times n)$ variance matrices, with W_t usually known and V_t known or unknown. Let Y_t be an $(r \times 1)$ vector observation on the time series Y_1, Y_2, \dots . Then the quadruple defines the model relating Y_t to the $(n \times 1)$ parameter vector θ_t at time t , and the θ_t sequence through time via the sequential specifications of distributions

$$(Y_t | \theta_t) \sim N[F_t' \theta_t, V_t], \quad (2.1)$$

$$(\theta_t | \theta_{t-1}) \sim N[G_t \theta_{t-1}, W_t]. \quad (2.2)$$

The above distributions are also conditional on D_{t-1} , the information set available prior to time t , but this conditioning is not explicitly recognised, only for the sake of notational simplicity.

We shall now consider the case of univariate NDLM. The above distributions can be represented alternatively by means of equations; so that the general univariate DLM is defined by:

$$\text{Observation : } Y_t = F_t' \theta_t + v_t, \quad v_t \sim N(0, V_t)$$

$$\text{System equation : } \theta_t = G_t \theta_{t-1} + w_t, \quad w_t \sim N(0, W_t)$$

$$\text{Initial prior : } (\theta_0 | D_0) \sim N[m_0, C_0],$$

where m_0, C_0 are prior moments.

The observation equation defines the sampling distribution for Y_t conditional on the parameter vector θ_t , together with observational error terms v_t . The system equation defines the dynamic structure of the model, with evolution error terms w_t . The observational and evolution error sequences are assumed to be independent and mutually independent, and are independent of $(\theta_0 | D_0)$.

The sets of v_t error represent simply a random perturbation in the measurement process which affects the observation Y_t but has no further influence on the series, whilst the sets of w_t error influence the development of the system into the future.

2.3.2 Updating Equations, Observational Variance V_t Known

The following standard results together with the forecast distributions are presented here and are referred to quite often in the thesis.

(i) Posterior at $t - 1$:

For some mean m_{t-1} and variance matrix C_{t-1} ,

$$(\theta_{t-1} | D_{t-1}) \sim N(m_{t-1}, C_{t-1})$$

(ii) Prior at t :

$$(\theta_t | D_{t-1}) \sim N(a_t, B_t)$$

where

$$a_t = G_t \Pi_{t-1} \quad \text{and} \quad R_t = G_t C_{t-1} G_t' + W_t.$$

(iii) One-step forecast:

$$(y_t | D_{t-1}) \sim N(f_t, Q_t),$$

where

$$f_t = F_t' a_t \quad \text{and} \quad Q_t = F_t' R_t F_t + V_t.$$

(iv) Posterior at t :

$$(B_t | D_t) \sim N(m_t, C_t),$$

with

$$m_t = a_t + A_t e_t \quad \text{and} \quad C_t = R_t - A_t Q_t A_t',$$

where

$A_t = R_t F_t' Q_t^{-1}$ and $e_t = y_t - f_t$, and, in particular, A_t is the regression matrix of B_t on y_t given D_{t-1} .

The forecast distributions for the series y_t and the state vector B_t are defined as follows. For each time t and

$k \geq 1$, the k -step ahead distributions for θ_{t+k} and y_{t+k} , given D_t , are given by:

(a) State distribution : $(\theta_{t+k} | D_t) \sim N[\hat{a}_t(k), B_t(k)]$,

(b) Forecast distribution : $(y_{t+k} | D_t) \sim N[\hat{f}_t(k), Q_t(k)]$,

with moments recursively defined by:

$$\hat{f}_t(k) = \hat{E}_{t+k}' \hat{a}_t(k)$$

and $Q_t(k) = \hat{E}_{t+k}' B_t(k) \hat{E}_{t+k} + V_{t+k}$

where $\hat{a}_t(k) = G_{t+k} \hat{a}_t(k-1)$,

and $R_t(k) = G_{t+k} R_t(k-1) G_{t+k}' + W_{t+k}$,

with starting values $\hat{a}_t(0) = \hat{m}_t$ and $R_t(0) = C_t$.

2.3.3 The Case of an Unknown Constant Observational Variance

$$v = \sigma^{-1}$$

2.3.3.1 Model Definition

For each t , the model is defined by

Observation equation : $y_t = F' \theta_t + v_t$, $v_t \sim N[0, V]$,

System equation : $\theta_t = G_t \theta_{t-1} + w_t$, $w_t \sim T_{n_{t-1}}[0, W_t]$

$$\text{Initial information : } (\theta_0 | D_0, \phi) \sim N(\underline{m}_0, \underline{C}_0 V)$$

$$(\phi | D_0) \sim G[n_0/2, d_0/2],$$

where $\phi = V^{-1}$, is the precision.

We are considering the case when V is constant but unknown (West and Harrison, 1989a). However, stochastic change in variance can be modelled though it is not discussed here.

2.3.3.2 Updating Equations

Given D_{t-1} , the information at time $(t-1)$ is

$$(i) \quad (\theta_{t-1} | D_{t-1}) \sim T_{n_{t-1}}(\underline{m}_{t-1}, \underline{C}_{t-1})$$

$$(\theta_{t-1} | D_{t-1}, \phi) \sim N(\underline{m}_{t-1}, \underline{C}_{t-1}/(\phi S_{t-1}))$$

$$(ii) \quad (\theta_t | D_{t-1}) \sim T_{n_{t-1}}(\underline{a}_t, \underline{R}_t),$$

$$(\theta_t | D_{t-1}, \phi) \sim N(\underline{a}_t, \underline{R}_t/(\phi S_{t-1}))$$

$$\text{where } \underline{a}_t = \underline{G}_t \underline{m}_{t-1} \text{ and } \underline{R}_t = \underline{G}_t \underline{C}_t \underline{G}_t' + \underline{W}_t$$

$$(iii) \quad (\phi | D_{t-1}) \sim G[n_{t-1}/2, d_{t-1}/2],$$

with $S_{t-1} = (E[\phi | D_{t-1}])^{-1} = d_{t-1}/n_{t-1}$ as a point estimate of V .

The updating recurrence relationships are given by

$$(\theta_t | D_t) \sim T_{n_t}(\bar{m}_t, \bar{C}_t)$$

$$(\theta_t | D_t \phi) \sim N(\bar{m}_t, \bar{C}_t / (\phi S_t))$$

$$(\phi | D_t) \sim G[n_t/2, d_t/2],$$

with $\bar{m}_t = \bar{a}_t + \bar{b}_t e_t$,

$$\bar{C}_t = (S_t/S_{t-1})[R_t - \bar{A}_t Q_t \bar{A}_t'] ,$$

$$n_t = n_{t-1} + 1, \quad d_t = d_{t-1} + S_{t-1} e_t^2 / Q_t, \quad \text{and} \quad S_t = d_t / n_t,$$

where $e_t = Y_t - f_t$, and $\bar{b}_t = R_t F_t / Q_t$.

The forecast distributions for the series Y_t and the state vector θ_t are defined as follows.

For each time t and for $k \geq 1$, the k -step ahead distributions for θ_{t+k} and Y_{t+k} , given D_t , are given by:

(a) State distribution: $(\theta_{t+k} | D_t) \sim T_{n_t}[\bar{a}_t(k), \bar{B}_t(k)]$

(b) Forecast distribution: $(Y_{t+k} | D_t) \sim T_{n_t}[f_t(k), Q_t(k)]$,

with moments as defined in subsection 2.3.2, with V_t replaced by the estimate $St-1$.

2.3.4 The DLM and Conditional Independence

Conditional independence, as coherently developed by Dawid (1979), offers a new framework for the development of statistical concepts as well as a useful tool in statistics. Thereafter, various authors (Lauritzen, J.Q. Smith etc.) have extended the ideas in the context of graphical models, influence diagrams, decision analysis etc. In particular, the properties of conditional independence help to understand how information is processed and transferred between uncertain quantities. In that context, conditional independence has proved useful in the study of DLMs and in developing new results in this thesis.

Conditional independence is briefly introduced here for future reference. Let W, X, Y, Z and U be random vectors. Then $X \perp\!\!\!\perp U \mid Z$ means that X is conditionally independent of U given Z . The following properties generally hold.

(i) $X \perp\!\!\!\perp Y \mid Z \Rightarrow Y \perp\!\!\!\perp X \mid Z$. This intuitively obvious symmetric property is useful from a practical point of view.

(ii) $X \perp\!\!\!\perp Y \mid Z \Leftrightarrow (X, Z) \perp\!\!\!\perp (Y, Z) \mid Z$.

(iii) If $X \perp\!\!\!\perp Y \mid Z$, and U is a function of X , then

(a) $U \perp\!\!\!\perp Y \mid Z$ and (b) $X \perp\!\!\!\perp Y \mid (Z, U)$.

(iv) If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (W, Y) \mid Z$.

The above properties are important and satisfy our intuition. For example (ii) could be explained as follows. If once Z is known, X conveys no information to Y , then X and Z together convey no information to Y and Z together.

One particular case where the above properties are useful is the discrete Markov-Chain, X_t . Let $Y_t = (X_{t+1}, X_{t+2}, \dots)$ denote the future at time t , and $Z_t = (\dots, X_{t-1}, X_t)$ denote the past at time t . Then the Markov property, in terms of conditional independence, is: for all t , $Y_t \perp\!\!\!\perp Z_t \mid X_t$, i.e. given the present, the future is independent of the past. This property is found to be most useful in the study of dynamic models.

In the Dynamic Linear Model where the state vector exhibits this Markovian property, the properties of conditional independence are most useful. This conditional structure in which the future and the past are conditionally independent given the present helps to develop recurrence relationships for routine learning and updating, forecasting, retrospective analysis and the calculation of diagnostics. This property will be illustrated by undirected graphs and used extensively in chapters four, five and six of the thesis.

2.4 Model Design and Specification

2.4.1 Introduction

The design of a model and the specification of various

characteristics of the model are obviously of paramount importance, and is briefly discussed in this section. Here the emphasis is on Time Series DLM's, TSDLM's, i.e. DLM's with F_t and G_t constant, hence denoted by F and G . The section deals first with parameterisation of the DLM, emphasising the need for parsimony and identifiability through the study of observability. Then, the specification of the F vector and the system matrix G is reviewed with the superposition principle as the main idea. Finally, the system variance vector W_t is discussed where the discounting principle plays a major role.

2.4.2 Observability

From the definition of the univariate TSDLM, the mean response function is defined by

$$\mu_{t+k} = E[Y_{t+k} | \theta_{t+k}] = F' \theta_{t+k}$$

and the forecast function is defined by

$$f_t(k) = E[\mu_{t+k} | D_t] = F' G^k \omega_t$$

Then it is easily shown that if $\mu_t = (\mu_t, \mu_{t+1}, \dots, \mu_{t+1-1})'$, we have $\mu_t = T \theta_t$, where T is the $n \times n$ matrix

$$T = \begin{pmatrix} F' \\ F' G \\ \vdots \\ F' G^{n-1} \end{pmatrix}$$

To precisely determine the state vector θ_t , T has to be non-singular so that $\theta_t = T^{-1} y_t$ is defined. Then the TSDLM $\{F, G, \dots, \dots\}$ is said to be observable if and only if the observability matrix T has full rank n .

If, however, T has rank $n - r$ for some r , ($1 \leq r < n$), then there exists a reparameterisation, based on linearly transforming the state vector, to an observable model of dimension $n - r$ that has the same mean response function. In this way, observability helps to achieve parsimony. By further taking into account characteristics of similarity and equivalence, we can then ensure as far as possible the identifiability of the parameter vector.

2.4.3 Superposition

In practice, models are required for real data series which very often exhibit a variety of features. Then the best approach, as adopted here, is to use simple DLMs to model these individual features and then to produce a global model for the series by collecting together in some way these simple component DLM's.

This can be effectively done by making use of the Principle of Superposition. It is simple, straightforward but most powerful from the model design point of view. It states that, under very general conditions, the linear combination of series generated by independent DLMs follows a DLM that is defined via the superposition of the corresponding model components. This principle is, of course,

dependent on additivity properties associated with linear models in general. Moreover, it is to be noted that the independence of models is not crucial for the principle to hold. But it is most important that the observational error terms, the V_t 's, are jointly normally distributed as well as the system error terms, the W_t 's.

In practical terms, the F_t 's for the components are easily defined by subvectors which together give the overall F_t for the global model. For the system matrices, G_t 's, we make use of the characteristics of Jordan block matrices (in particular their eigenvalue configuration) to bring together the system matrices of the components DLMs. The system error matrices, the W_t 's, would be developed in a similar way. In fact both would be block diagonal Jordan matrices.

2.4.4 Discounting

The idea of discounting has been used previously by classical statisticians. In particular, Brown (1962) developed the idea of discounting information; but then, he used only one discount factor for the global time series model. By 1965, Harrison was the first to introduce the idea of using different discount factors for different components. This is an important idea used in developing the system error matrices, W_t 's.

The key idea is based on the fact that the system error term models a loss of information or a decay of information between

observations. Using previous notation, we have

$$\begin{aligned} R_t &= V(\theta_t | D_{t-1}) \\ &= G_t C_{t-1} G_t' + W_t \\ &= P_t + W_t, \text{ where } P_t \text{ is obviously defined.} \end{aligned}$$

In the ideal case of no system error, R_t and P_t would be equal; in practice, W_t brings in uncertainty, so that we can view P_t as a discounted R_t , with a discount factor δ , $0 < \delta < 1$. Thus we have $R_t = P_t/\delta$ which implies that $W_t = P_t(1 - \delta)/\delta$.

Bearing in mind the principle of superposition, it follows in a fairly straightforward way that, for each component, the system error matrix may have a different discount factor. This means that the decay of information over time for each component model is being modelled at a different rate. And this gives obviously much flexibility to the model as a whole. It allows it to be responsive to various sources of uncertainty so that the global model is more robust than otherwise it would have been.

It is to be noted that when Harrison and Stevens (1976) presented the new conceptual framework of Bayesian forecasting and dynamic models, they then produced a full covariance matrix for the system error component. Not only was it rather cumbersome to work with for practitioners, but the approach using discounting overcomes various other shortcomings. It removes ambiguity and also removes the lack of invariance to units in which independent variables are measured. Further, it provides a unified simple way of specifying durability of quantified model form.

2.5 Dealing with Model Uncertainty

2.5.1 Introduction

Some general ideas related to model uncertainty were introduced in section 2.3.1 in the context of the discussion on dynamic models. This section deals briefly with more specific aspects of model uncertainty together with a discussion of the various approaches developed to handle that uncertainty.

Model uncertainty manifests itself in the inadequacy of the model as revealed mainly in the forecast performance and in the production of other measures like the mean etc. To handle that inadequacy we need to be precise as to what is its nature. Inadequacy in a model can be categorised along two broad lines:

- (i) inadequacy in the definition of the parameters and other structural features like the components of the quadruple $\{F_t, G_t, V_t, W_t\}$, without the model being found globally inadequate;
- (ii) the model is found globally inadequate, so that alternatives need to be found.

The ideas and techniques of filtering, intervention and monitoring are used to handle model inadequacy as defined in (i) above, whilst multiprocess models are used to handle cases which fall in category (ii).

2.5.2 Filtering and Smoothing

There is often the need for inferences about the state of a process under study in the past, at times $t - k$ for $k > 0$. This is the type of procedure used in process monitoring and control. The act of using recent data to revise inferences about previous values of the state vector is called filtering; whilst the retrospective estimation of the historical development of a time series mean response function is called smoothing the series. The whole idea consists of obtaining the k -step filtered distribution of the state vector at that time, i.e. the distribution of $(\theta_{t-k} | D_t)$, for $k \geq 1$ and any fixed t ; then smoothing of the time series involves using the filtered distribution for the mean response function μ_t .

These techniques are well known standard ones; the results, analogous to the k -step ahead forecast distribution in the case of the state vector, are obtained using Bayes' Theorem. Later in Chapter Four it will be shown how the full retrospective probability distribution may be derived as a particular case of incorporating information.

2.5.3 Intervention

Intervention is basically the interaction between the forecaster and the model; this characterises forecasting systems which include forecasters as integral components. The unique 'true' model, and even 'objective' model, just does not make sense.

Interventions can be broadly categorised in either feed-

forward or feed-back. As the names indicate, the former is anticipatory in nature whilst the latter is corrective, responding to events that had not been adequately catered for. The latter is carried out following model assessment carried out by a given monitoring mechanism and will be dealt with briefly at the end of this section. We now deal mainly with feed-forward interventions.

There are some feed forward interventions which do not affect the dimension of the model and there are some which bring amendments to the model dimension. Those in the first category can be summarised as follows:

- (i) Treating an observation, Y_t , as an outlier and therefore ignoring it. This happens when there are unusual events which produce a single discrepant observation. In such cases, it should not be used in updating the model for forecasting.
- (ii) Conditions in the environment of a given process may affect the time series. For example, the introduction of competitor activity in a consumer market will bring uncertainty about the future, reflected by increased uncertainties about some or all of the existing parameters. This leads therefore to additional noise in the system, so that the system variance would then be increased.
- (iii) There is then the case of arbitrary subjective intervention where generally the prior moments of the state vector θ_t are changed to new values, in anticipation of changes in the series.

In all these cases, the interventions can be defined in such a way that they can be incorporated in the form of DLM, thus making post intervention analysis very simple. Details are omitted here.

Finally, the case when intervention brings change to the dimension of the model arises when it is of interest to isolate the effects of an intervention. Then extra parameters are provided. For example if an increase in level is suspected, the prior estimate is not merely increased, but the parameter space is modified from say $\{\theta_t\}$ to $\{\theta_{t1}, \theta_{t2}\}$ where $\theta_t = \theta_{t1} + \theta_{t2}$.

The idea of intervention is, of course, to cater for model inadequacy so that, thereafter, the model is improved and that leads to more rapid adaptation in the future. The ability to incorporate such interventions appropriately in the model as mentioned above helps furthermore to render the models self correcting if the inadequacies can be identified and accordingly defined at points of suspected change.

Feed-back intervention is carried out following model breakdown detected by a monitoring mechanism. The intervention deals mainly with:

- (i) the omission of an observation, identified as an outlier by the monitoring mechanism;
- (ii) change regarding the parameters, again following model breakdown.

Such feed-back intervention can be incorporated in a self-correcting mechanism.

2.5.4 Monitoring

Model assessment and monitoring is a key concept in a scientific approach to modelling. It provides tools to detect model failure, to diagnose it and hence it creates conditions to improve the model performance: it helps, in other words, the learning process.

In the DLM framework, continual sequential monitoring is carried out by detecting deteriorations in predictive performance that are consistent with some form of model breakdown. This implies the need to characterise the type of model breakdown and the need to measure predictive performance in some way.

The characterization of model breakdown means simply that the model performance is assessed relative to that obtained using one or more alternative models. These alternative models would be designed to cater for outlying observations, or changes in parameters etc.

Regarding the measurement of predictive performance, the Bayes' Factor is used to measure relative predictive performance of two given models, with the emphasis being on measuring consistency of each observation with the corresponding one-step ahead forecast distribution. In other words, the focus lies on the local performance of the model, not on the historical performance. For

some form of intervention is considered desirable to the extent that the current and most recent observations do not accord with the model.

The Bayes Factor for model M_0 versus model M_1 based on the observed value of Y_t is defined as

$$H_t = P_0(Y_t | D_{t-1}) / P_1(Y_t | D_{t-1}) ,$$

where M_0 would be the routine, standard model and M_1 the alternative model. The overall Bayes' Factor, $H_t(k)$, based on k consecutive observations is equally useful. Making use of H_t and $H_t(k)$, we can then detect model breakdown using the criteria given by Jeffreys (1961): a log Bayes factor of 1 indicates evidence in favour of model 0, a value of 2 or more indicating the evidence to be strong.

Later, in Chapter 8, these ideas are used and elaborated in the context of developing an appropriate model monitoring mechanism.

2.5.5 Multi-Process Models

Wherever the global inadequacy of any single DLM is acknowledged, there is then the need for alternative models as highlighted in the discussion on model monitoring. A multi-process model effectively caters for such inadequacy: it is a combination of several DLMs. More precisely, it is a mixture of several DLMs. In the case of Multiprocess Models Class II (defined later), then there would be an explosion in the number of likelihood functions as the

process unfolds itself; this is usually handled by means of an appropriate approximating method as mentioned later on.

We introduce some notation to help to make the ideas clear. Consider DLMs M_t which depend on a parameter α which is subject to uncertainty. It may be, for example, discount factor etc; we may write then $M_t = M_t(\alpha)$, and let A denote the set of possible values of α . Then, either of the following possibilities may occur:

- I. For some $\alpha_0 \in A$, $M_t(\alpha_0)$ holds for all t . In such cases, there is uncertainty with regard to the value of α , so that, usually, the different models would be structurally similar with different estimates of α_0 .
- II. For some sequence of values $\alpha_t \in A$, ($t = 1, 2, \dots$), $M_t(\alpha_t)$ holds at time t . In these cases, it is recognised that different models are appropriate at different times, thus conveying in practical terms the idea of dynamic models discussed earlier on.

The above classes of models are known respectively as Multi-Process Models Class I and Multi-Process Models Class II. In the first category, we may have, for example, models with different values of the discount factor, whilst being structurally similar. In the second category, we may have very different structural models like an outlier model, a growth model etc.

The model selection probabilities are obviously very

pertinent; when they are known, the problem is simplified. Otherwise there are learning procedures to help us to update any prior beliefs regarding them. Very often, fixed model selection probabilities, but unknown, would be the case.

In both cases, there is a need to evaluate the mixture of the models, so that both are constrained as a result from the practical point of view given the complexity involved. It is to be noted that only an approximate density can be obtained, using a collapsing procedure especially in the case of Multi-Process Models Class II. The Kullback-Leibler directed divergence is used and minimised to give the approximating mixture. If such a collapsing procedure is not used, there would be an explosion of likelihoods and the problem would then be unsolvable!

The Multi-Process Models Class II are certainly very realistic and provide a most powerful tool to modelling particularly complex processes. It seems it is a rather underutilised tool though, understandably, practitioners may find it difficult to handle.

2.6 Extensions to the NDLM

So far, we have been discussing normal DLMs. There have been major developments whereby the dynamic models need not be linear nor normal.

Non-linear Dynamic Models have been developed whereby the structure of the DLM has been maintained. The idea is based on

three approaches:

- (i) Various linearisation techniques are used, based essentially on the use of linear approximation to non-linearities.
- (ii) Some models have non-linearities due to the appearance of constant, but unknown, parameters α in one or more of the components. They are usually analysed within the framework of multi-process models, Class I mentioned in Section 2.5.5.
- (iii) Refined techniques of numerical integration based on Gaussian quadrature are used to approximate the posterior distributions.

In the cases of non-linear models discussed above, the normality of the DLMs is usually preserved. Moreover, the DLMs have been extended to cover non-normal problems, in particular the exponential family models. In relation to classical static modelling, corresponding dynamic generalised linear models have been developed.

Without going into any detailed discussion, it is perhaps sufficient to note that a conjugate analysis of the exponential family of models in terms of the natural parameter, η_t has been first developed. Then, the structure of the DLM has been maintained, whilst making use of linear Bayesian estimation methods where necessary to obtain the posterior moments.

Finally, the theoretical framework for multivariate DLMs has been developed so that the DLM provides a comprehensive framework for modelling, as in West and Harrison (1989a, Chapter 15).

2.7 Final Comments on DLMs

Among the latest developments regarding the DLMs, research in the field of multivariate DLMs has been particularly prominent. The research has focussed mainly on multivariate normal DLMs with the work of Quintana (1987), Quintana and West (1987, 1988), Barbosa (1989) and Barbosa and Harrison (1989). Further developments dealing with non-normal and non-linear models have been achieved, whereby ideas of conditional independence and graphical models have been used as in Queen and Smith (1989, 1990).

In the area of diagnostics, Harrison and West (1990) have studied the impact of leaving out one observation on past state parameters. In this thesis, techniques and recurrence relationships regarding the deletion of more than one observation on the one hand and regarding the incorporation of information on the other hand have been developed. Properties of conditional independence are found to be most useful.

2.8 Bayesian Forecasting and the Classical Approach

Whilst the thrust of the thesis is within the framework of Bayesian Forecasting and Dynamic Models, it is appropriate to place the development of such a framework within the global evolution of the theory and applications of time series analysis and forecasting.

The study of time series analysis and forecasting has been traditionally dominated by broadly two approaches: the structural approach and the approach linked to autoregressive and moving average

(ARMA) models. The development of the latter has culminated in the quite well known Box and Jenkins approach where the emphasis on mathematics is very pronounced.

Within the structural approach, a major development took place when the idea of discounting in the form of Exponentially Weighted Regression was developed by Brown (1962); he introduced the notion of information decay and, in practical terms, the discounting principle. But then, as noted earlier on in this chapter, only one discount factor was used for all components of the time series. Harrison (1965) introduced the idea of using multiple discount factors; he used different ones for the trend and the seasonal component. At about the same time (1964) he developed a monitoring mechanism using the cusums for forecasting systems. Also, the Holt-Winters forecasting procedure, i.e. the generalisation of exponential smoothing to deal with time series containing trend and seasonal variation, was then developed.

By late 1960s and early 1970s, a qualitative change of fundamental importance took place when Harrison and Stevens (1971, 1976) developed a Bayesian approach to forecasting and time series analysis: key ideas of dynamic modelling, multi-process models and applied Bayesian methodology were introduced. Since then, major developments have taken place.

Meanwhile, a non-Bayesian approach has been developed whilst using dynamic models by, amongst others, Harvey (1981) and Young (1984). It is quite interesting to note that the dualistic approach, Bayesian and non-Bayesian, survives among those using the structural approach.

CHAPTER THREE

GENERAL RESULTS FOR NORMAL DYNAMIC MODELS

3.1 Introduction

The first part of this thesis starts with this chapter; it deals with handling of information in normal dynamic models, in particular with the incorporation of information and with the deletion of suspect information/unusual values. It covers Chapter 3 to Chapter 6, whilst the second part of the thesis, which is covered in Chapter 7 to Chapter 10, deals with model monitoring, the emphasis being on the detection of unusual values/outliers.

In this Chapter, we deal specifically with general results for normal dynamic models, making use of results generally applicable to normal models. Strong conditions, like conditional independence properties, are not required for these results. Moreover, in subsequent chapters, results for normal dynamic models, which make use of conditional independence properties of normal models developed in Chapter 4, are dealt with.

We now introduce some notation slightly different to the one used in Chapter 2, and similar to the one used in Harrison and West (1991), regarding, in particular, the filtered distribution; it would help to bring clarity to the presentation of the new ideas and techniques. Using some of the same symbols used in Section 2.3 of Chapter 2, let D_n denote the information set $\{D_0, (Y_t ; t = 1, \dots, n)\}$; then the state vectors will be normally distributed

with

$$\theta_t | \phi, D_n \sim N(a_{n,t}; R_{n,t}/\phi) \quad (t = 1, \dots, n) \quad (3.1)$$

The notational implications for the DLM are clear; for example, we have

$$(i) \quad \theta_t | D_t, \phi \sim N(a_{t,t}; R_{t,t}/\phi) \text{ instead of } N(m_t; C_t/\phi).$$

$$(ii) \quad \theta_t | D_{t-1}, \phi \sim N(a_{t-1,t}; R_{t-1,t}/\phi) \text{ instead of } N(a_t; R_t/\phi).$$

3.2 Standard Normal Results for Incorporating Information

Let the $h \times 1$ vector U be any set of observations, c a known matrix, and Z any set of observations and states, not necessarily including the states corresponding to the observational times. Moreover, U may also represent subjective information and external forecasts which are modelled as in West and Harrison (1989), so that the results obtained in this section and chapter are widely applicable. Also, in this chapter and subsequent chapters, statements, such as ' A is the regression matrix of Z on U ', are to be understood as being conditional upon a joint normal distribution with known variance.

Let the following be proper distributions

$$(U | Z = z, \phi) \sim N(c'z; \sum_{u|z} / \phi), \quad (3.2)$$

$$(Z | U = u, \phi) \sim N(\mu_z | u; \sum_{z|u} / \phi),$$

$$(Z | \phi) \sim N(\mu_Z; \Sigma_Z / \phi) .$$

With A being the regression matrix of Z on U , the following expressions are defined:

- (i) $e = u - c'\mu_Z$, i.e. the error in the estimate provided by the model defined by (3.2), with

$$E(U | Z = z, \phi) = c'\mu_Z$$

- (ii) from standard multivariate joint normal distribution results, (West and Harrison, 1989a, Chapter 16)

$$(a) \quad \Sigma_U = \Sigma_{U|Z} + c' \Sigma_Z c \quad (3.3)$$

$$(b) \quad A = A_{Z,U} = \Sigma_Z c \Sigma_U^{-1}$$

Then, standard normal results for incorporating the information $U = u$ are given by (West and Harrison, 1989a)

$$\begin{aligned} \mu_{Z|U} &= \mu_Z + A(u - c'\mu_Z) \\ &= \mu_Z + A e \end{aligned} \quad (3.4)$$

$$\begin{aligned} \Sigma_{Z|U} &= \Sigma_Z - A \Sigma_U A' \\ &= \Sigma_Z - A \Sigma_U \Sigma_U^{-1} c' \Sigma_Z, \text{ using definition of } A \end{aligned}$$

$$= \sum_z - A c' \sum_z \quad (3.5)$$

In parallel to the other results, we shall now develop dual expressions for eliminating information $U = u$. These expressions would obviously be required for the calculation of deletion diagnostics and for discarding suspect data sets. It is to be stressed that these expressions are valid for normal models with no other conditions attached, which mean they are widely applicable.

We shall need the following standard results of multivariate normal distribution which have been derived by making use of the multivariate Bayes' Theorem (West and Harrison, 1989a, Chapter 16)

$$\sum_z^{-1} | u = \sum_z^{-1} + c \sum_u^{-1} | z c' \quad (3.6)$$

$$\sum_z^{-1} | u \mu_z | u = \sum_z^{-1} \mu_z + c \sum_u^{-1} | z u \quad (3.7)$$

3.3 Leverage of an observation on the States

Before obtaining the required expressions, we shall first obtain an expression for the leverage of U on Z , which is measured by A , the regression matrix of Z on U . From (3.7), we may get an intuitive feel of the influence or leverage of an observation in parameter estimation. If Z is a vector of states,

then the posterior mean is given by

$$\mu_{z|u} = \sum_{z|u} \sum_z^{-1} \mu_z + \sum_{z|u} c \sum_{u|z}^{-1} u,$$

which is the weighted average of the prior estimate of the mean of z , given by μ_z , and the observation u . The weight $\sum_{z|u} c \sum_{u|z}^{-1}$ assigned to u gives intuitively a measure of leverage of u in the estimation of the mean of z .

This result will be proved formally in the following lemma; the result is not only important on its own merit but it is important to derive recurrent relationships for the moments of the distribution of z with u deleted as well as for the jackknifed residual, particularly important to obtain deletion diagnostics along the approach adopted by Harrison and West (1991).

Lemma : Given $U|z$ and $Z|u$, the leverage of U on Z

as measured by A , is calculable as

$$A = A_{z,u} = \sum_{z|u} c \sum_{u|z}^{-1} \quad (3.8)$$

Proof:

$$\begin{aligned} A &= \sum_z c \sum_{u|z}^{-1}, \text{ by definition} \\ &= \sum_z c (c' \sum_z c + \sum_{u|z})^{-1}, \text{ using (3.3)} \end{aligned} \quad (3.9)$$

Now, post multiplying (3.6) by A , we have

$$\begin{aligned}\Sigma_{z|u}^{-1} A &= \Sigma_z^{-1} A + c \Sigma_{u|z}^{-1} c' A \\ &= c(c' \Sigma_z c + \Sigma_{u|z})^{-1} + c \Sigma_{u|z}^{-1} c' \Sigma_z c(c' \Sigma_z c + \Sigma_{u|z})^{-1},\end{aligned}$$

using (3.9).

$$\begin{aligned}&= c[I + \Sigma_{u|z}^{-1} c' \Sigma_z c](c' \Sigma_z c + \Sigma_{u|z})^{-1} \\ &= c \Sigma_{u|z}^{-1} [\Sigma_{u|z} + c' \Sigma_z c](c' \Sigma_z c + \Sigma_{u|z})^{-1} \\ &= c \Sigma_{u|z}^{-1}\end{aligned}$$

$$\text{Hence, } A = \Sigma_{z|u} c \Sigma_{u|z}^{-1} \cdot \square$$

3.4 Deletion of Information

3.4.1 The Case of Known Variance

To obtain the required recurrence relationships we introduce some notation; let

$$d = u - c' \mu_{z|u}; \quad (3.10a)$$

$$q = \Sigma_{u|z} - c' \Sigma_{z|u} c; \quad (3.10b)$$

$$B = \sum_{z|u} c \, q^{-1} . \quad (3.10c)$$

Theorem 3.1

- (i) The moments of the distribution of Z with u deleted are

$$\mu_z = \mu_{z|u} - B d , \quad (3.11)$$

$$\sum_z = \sum_{z|u} + B c' \sum_{z|u} , \quad (3.12)$$

- (ii) The jackknifed residual e is calculable as

$$e = \sum_u \sum_{u|z}^1 d = \sum_{u|z} q^{-1} d \quad (3.13)$$

Proof:

- (i) $A = \sum_{z|u} c \sum_{u|z}^1$, from the lemma
- $$= \sum_{z|u} c \, q^{-1} q \sum_{u|z}^1$$
- $$= B \, q \sum_{u|z}^1 , \text{ using (3.10c)}$$
- $$= B (\sum_{u|z} - c' \sum_{u|z} c) \sum_{u|z}^1 , \text{ using (3.10b)}$$
- $$= B - B c' \sum_{z|u} c \sum_{u|z}^1$$
- $$= B - B c' A , \text{ using lemma.}$$

So that

$$A = B(I - C'A) ,$$

$$B = A(I - C'A)^{-1} .$$

Now,

$$\begin{aligned}(I - AC') A &= A - AC'A \\ &= A(I - C'A)\end{aligned}$$

So that

$$A(I - C'A)^{-1} = (I - AC')^{-1} A$$

and hence

$$B = (I - AC')^{-1} A . \quad (3.14)$$

Now, we shall establish another intermediate result, $B = \sum_z C \sum_u^{-1} z$.

Premultiplying (3.6) by $\sum_z | u$ gives

$$I = \sum_z | u \sum_z^{-1} + \sum_z | u C \sum_u^{-1} z C' .$$

and now post multiplying by $\sum_z C \sum_u^{-1} z$ gives

$$\begin{aligned}\sum_z C \sum_u^{-1} z &= \sum_z | u C \sum_u^{-1} z + \sum_z | u C \sum_u^{-1} z C' \sum_z C \sum_u^{-1} z \\ &= A + A C' \sum_z C \sum_u^{-1} z .\end{aligned}$$

Hence

$$(I - AC') \sum_z C \sum_u^{-1} z = A$$

So that

$$\begin{aligned}\sum_z C \sum_u^{-1} z &= (I - AC')^{-1} A \\ &= B , \text{ using (3.14)}\end{aligned}$$

Then, premultiplying (3.7) by \sum_z gives

$$\begin{aligned}\sum_z \sum_z^{-1} u \mu_z | u &= \mu_z + \sum_z c \sum_u^{-1} u , \\ &= \mu_z + Bu ,\end{aligned}$$

so that

$$\mu_z - \sum_z \sum_z^{-1} u \mu_z | u = Bu \quad (3.15)$$

Now, premultiplying (3.6) by \sum_z gives

$$\begin{aligned}\sum_z \sum_z^{-1} u &= I + \sum_z c \sum_u^{-1} c , \\ &= I + Bc ,\end{aligned} \quad (3.16)$$

so that

$$\sum_z \sum_z^{-1} u \mu_z | u = \mu_z | u + Bc \mu_z | u \quad (3.17)$$

Adding (3.15) to (3.17) gives

$$\mu_z + \sum_z \sum_z^{-1} u \mu_z | u - \mu_z | u + \sum_z \sum_z^{-1} u \mu_z | u - Bu + Bc \mu_z | u ,$$

so that

$$\mu_z = \mu_z | u - B(u - c \mu_z | u) ,$$

hence,

$$\mu_z = \mu_z | u - Bd , \text{ proving thus (3.11)}$$

Now, proving (3.12) is quite straightforward. Postmultiplying (3.16) by $\sum_z | u$ gives

$$\sum_z = \sum_z | u + Bc' \sum_z | u, \text{ thus establishing (3.12) .}$$

(ii) $e = u - c' \mu_z$, by definition

$$= u - c' \mu_z | u + c' (\mu_z | u - \mu_z)$$

$$= d + c' Ae, \text{ using (3.10a) and (3.4)}$$

so that

$$d = (I - c' A) e$$

$$= (I - c' \sum_z c \sum_u^{-1}) e, \text{ by definition of } A$$

$$= (\sum_u - c' \sum_z c) \sum_u^{-1} e$$

$$= \sum_u | z \sum_u^{-1} e, \text{ using (3.3)}$$

hence

$$e = \sum_u \sum_u^{-1} | z d, \text{ proving the first part of (3.13) .}$$

Now, from definition of d in (3.10a),

$$d = u - c' \mu_z | u$$

$$= u - c' \mu_z - c' B d, \text{ using (3.11)}$$

$$= u - c' \mu_z - c' \sum_z | u c q^{-1} d, \text{ using (3.10c)}$$

so that

$$u - c' \mu_z = d + c' \sum_z | u \ c q^{-1} d ,$$

$$e = (q + c' \sum_z | u \ c) q^{-1} d , \text{ by definition of } e$$

$$= (\sum_u | z - c' \sum_z | u \ c + c' \sum_z | u \ c) q^{-1} d, \text{ from (3.10b)}$$

$$= \sum_u | z \ q^{-1} d \quad \square$$

3.4.2 The Case of Unknown Variance

For the unknown variance case, we need the results of the usual conjugate analysis to learn on the variance or precision, carried out using the normal-gamma distribution as in West and Harrison (1989a; Chapter 16). Bearing in mind the equivalence between the Gamma distribution and the χ^2 distribution, i.e. $\phi \sim \text{Ga}[n/2, d/2]$ is equivalent to $d\phi \sim \chi^2_n$, n being a positive integer, the prior distribution for ϕ is defined by

$$vs\phi \sim \chi^2_v , \tag{3.18}$$

where v, s are known constants and v is the number of degrees of freedom. Then given $U = u$, the posterior distribution of ϕ is defined by

$$v_u s_u \phi | u \sim \chi^2_{v_u} , \text{ with}$$

$$v_u = v + h, \quad (3.19)$$

and

$$v_u s_u = v s + e' \sum_u^{-1} e, \quad (3.20)$$

where h = dimension of the vector U .

Theorem 3.2

Given the distribution $v_u s_u \mid u \sim \chi^2_{v_u}$, the distribution of θ with the information $U = u$ deleted may be obtained by calculating

$$v = v_u - h \text{ and}$$

$$v s = v_u s_u - d' q^{-1} d \quad (3.21)$$

Proof:

Making use of the additive property of the χ^2 distribution, it follows in a straightforward way that the distribution of θ with information $U = u$ deleted is given by a χ^2 distribution with number of degrees of freedom $v = v_u - h$, using (3.19) and $v s = v_u s_u - e' \sum_u^{-1} e$, using (3.20).

But from (3.13) of Theorem 3.1, we have

$$\sum_u \sum_{u|z}^{-1} d = \sum_{u|z} q^{-1} d ,$$

so that

$$\sum_{u|z}^{-1} \sum_{u|z}^{-1} d = q^{-1} d , \text{ premultiplying } \sum_{u|z}^{-1} \quad (3.22)$$

Hence

$$\begin{aligned} e' \sum_u^{-1} e &= d' \sum_{u|z}^{-1} \sum_u \cdot \sum_u^{-1} \cdot \sum_{u|z}^{-1} d , \text{ using again} \\ (3.13) \quad &= d' \sum_{u|z}^{-1} \sum_{u|z}^{-1} d \\ &= d' q^{-1} d , \text{ using (3.22) . } \quad \square \end{aligned}$$

The theorem 3.2 is of great practical significance. It helps to make the usual computations shorter than it would have been otherwise. This result of theorem 3.2 is equally important when further results are developed in Chapter 5.

3.5 The Special Case of the Dynamic Linear Model

The Dynamic Linear Model, as defined and discussed in Section 2.4 of Chapter 2, is obviously a special case of the model defined in Section 3.2 of the present chapter and which has been discussed so far. We can take $U = Y_t$, $Z = \theta_t$, and $c = F_t$. Bearing in mind the new notation introduced as per (3.1) in Section 3.1, and conditioning everything on D_{t-1} , the correspondence is

given by the following:

$$Y_t | \theta_t, \phi \sim N(F_t' \theta_t; V_t/\phi),$$

$$\theta_t | D_t, \phi \sim N(a_{t,t}; R_{t,t}/\phi),$$

$$\theta_t | D_{t-1}, \phi \sim N(a_{t-1,t}; R_{t-1}/\phi),$$

$$v_t a_t \phi | D_t \sim \chi_{v_t}^2,$$

$$a_{t-1,t} = G_t a_{t-1,t-1},$$

$$R_{t-1,t} = G_t R_{t-1,t-1} G_t' + W_t,$$

$$e_t = Y_t - F_t' a_{t-1,t},$$

$$\text{Var}(Y_t | D_{t-1}, \phi) = Q_t/\phi, \text{ where } Q_t = F_t' R_{t-1,t} F_t + V_t.$$

It is to be noted that the regression matrix of θ_t on Y_t given D_{t-1} is written A_t , where $A_t = R_{t-1,t} F_t' Q_t^{-1}$. The above model is obviously defined in the case when variance is known; then we take $\phi = 1$, so that there is no learning on ϕ .

CHAPTER FOUR

CONDITIONAL INDEPENDENCE FOR NORMAL DISTRIBUTIONS

4.1 Introduction

After having studied the incorporation and deletion of information in a general setting in Chapter 3, we now develop powerful results for normal dynamic models by making use of the properties of conditional independence. Procedures for data incorporation with particular application to dynamic linear models are discussed; those for data deletion are discussed in Chapter 5.

It is appropriate to recall the notation used for conditional independence. If X and U are conditionally independent given Z , we write $X \perp\!\!\!\perp U \mid Z$.

4.2 Important Results

In the following theorems, let X , Z and U be normal random vectors and write the regression matrices of X on Z and Z on U as $A_{X,Z}$ and $A_{Z,U}$ respectively so that

$$\begin{pmatrix} X \\ Z \\ U \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_X \\ \mu_Z \\ \mu_U \end{pmatrix} ; \begin{bmatrix} \Sigma_X & A_{X,Z}\Sigma_Z & A_{X,U}\Sigma_U \\ \cdot & \Sigma_Z & A_{Z,U}\Sigma_U \\ \cdot & \cdot & \Sigma_U \end{bmatrix} \right)$$

Let X and U be conditionally independent given Z .

Theorem 4.1

The regression matrix $A_{X,U}$ of X on U is the product of the regression matrices of X on Z and of Z on U , so that

$$A_{X,U} = A_{X,Z} A_{Z,U} , \quad (4.1)$$

$$\text{Cov}(X,U) = A_{X,Z} A_{Z,U} \sum_U \quad (4.2)$$

Proof:

Because of normality,

$$X \perp\!\!\!\perp U \mid Z \iff 0 = \text{Cov}(X, U \mid Z) .$$

So, using normal conditional results,

$$0 = \text{Cov}(X, U) - \text{Cov}(X, Z) \sum_Z^{-1} \text{Cov}(Z, U)$$

Therefore, $\text{Cov}(X, U) = A_{X,Z} \text{Cov}(Z, U)$

$$\begin{aligned} \text{Now, } A_{X,U} &= \text{Cov}(X, U) \sum_U^{-1} \\ &= A_{X,Z} \text{Cov}(Z, U) \sum_U^{-1} \\ &= A_{X,Z} A_{Z,U} . \end{aligned}$$

The result $\text{Cov}(X, U) = A_{X,Z} A_{Z,U} \sum_U$ is then obvious. \square

Result (4.1) can now be generalised in a straightforward

way, as spelt out in the following corollary.

Corollary 4.1

IF X_1, X_2, \dots, X_n are random vectors such that for all i and for all $1 < j < k < n$, $X_i \perp\!\!\!\perp X_{i+k} \mid X_{i+j}$, with $A_{i,i+1}$ as the regression matrix of X_i on X_{i+1} , then the regression matrix of X_1 on X_n is

$$A_{1,n} = \prod_{i=1}^{n-1} A_{i,i+1}$$

Proof:

The proof is by induction.

By putting $X = X_1$, $Z = X_2$, $U = X_3$, then $X_1 \perp\!\!\!\perp X_3 \mid X_2$, so that result (4.1) holds giving

$$A_{1,3} = A_{1,2} A_{2,3}.$$

Thus the result holds for $n = 3$. Suppose now that the result is true for $n - 1$. Then we have

$$A_{1,n-1} = \prod_{i=1}^{n-2} A_{i,i+1}$$

But $X_1 \perp\!\!\!\perp X_n \mid X_{n-1}$, so that again result (4.1) holds so that

$$\begin{aligned}
 A_{1,n} &= A_{1,n-1} \cdot A_{n-1,n} \\
 &= \left(\prod_{i=1}^{n-2} A_{i,i+1} \right) \cdot A_{n-1,n} \\
 &= \prod_{i=1}^{n-1} A_{i,i+1} \quad . \quad \square
 \end{aligned}$$

Theorem 4.2

Define the conditional distributions $X|U \sim N(\mu_{X|U}; \Sigma_{X|U})$ and $Z|U \sim N(\mu_{Z|U}; \Sigma_{Z|U})$. Then, given $U = u$, and $X \perp\!\!\!\perp U|Z$

- (i) X and Z are jointly normal, with

$$\mu_{X|U} = \mu_X + A_{X,Z}(\mu_{Z|U} - \mu_Z) \quad (4.4)$$

$$\Sigma_{X|U} = \Sigma_X + A_{X,Z}(\Sigma_{Z|U} - \Sigma_Z) A'_{X,Z} \quad (4.5)$$

- (ii) Given u , the regression matrix $A_{(X,Z)|U}$ of X on Z remains as $A_{X,Z}$ with

$$\text{Cov}(X, Z|U) = A_{X,Z} \Sigma_{Z|U} \quad (4.6)$$

- (iii) But the regression matrix $A_{(Z,X)|U}$ changes so that

$$A_{(Z,X)|U} = \Sigma_{Z|U} A'_{X,Z} \Sigma_{X|U}^{-1} \quad (4.7)$$

- (iv) The regression matrix of X on U may be written

$$A_{X,U} = A_{X,Z} A_{Z,U} = \Sigma_{X|U} A'_{(Z,X)|U} \Sigma_{Z|U}^{-1} A_{Z,U} \quad (4.8)$$

Proof:

The proof follows using standard conditional normal results mentioned in Section 3.2 of Chapter 3, i.e. results (3.4) and (3.5).

(i) From multivariate normal theory, it is clear that given $U = u$, X , Z are jointly normal.

From (3.4), we have

$$\mu_{Z|U} = \mu_Z + A_{Z,U}(u - \mu_U),$$

so that $A_{Z,U}(u - \mu_U) = \mu_{Z|U} - \mu_Z$.

Then, $\mu_{X|U} = \mu_X + A_{X,U}(u - \mu_U)$, using again (3.4)

$$= \mu_X + A_{X,Z}A_{Z,U}(u - \mu_U)$$

$$= \mu_X + A_{X,Z}(\mu_{Z|U} - \mu_Z).$$

Also, $\Sigma_{Z|U} = \Sigma_Z - A_{Z,U} \Sigma_U^{-1} A'_{Z,U}$, using standard conditional normal results, so that

$$A_{Z,U} \Sigma_U^{-1} A'_{Z,U} = \Sigma_Z - \Sigma_{Z|U}.$$

Then, using again standard conditional normal results,

$$\begin{aligned} \Sigma_{X|U} &= \Sigma_X - A_{X,U} \Sigma_U^{-1} A'_{X,U} \\ &= \Sigma_X - A_{X,Z}A_{Z,U} \Sigma_U^{-1} A'_{Z,U}A'_{X,Z} \\ &= \Sigma_X - A_{X,Z}(\Sigma_Z - \Sigma_{Z|U})A'_{X,Z} \\ &= \Sigma_X + A_{X,Z}(\Sigma_{Z|U} - \Sigma_Z)A'_{X,Z}. \end{aligned}$$

$$(ii) \quad \text{Cov}(x, z | u) = \text{Cov}(x, z) - \text{Cov}(x, u) [\text{Var}(u)]^{-1} \text{Cov}(u, z)$$

$$= A_{x,z} \Sigma_z - A_{x,u} \Sigma_u \Sigma_u^{-1} [\text{Cov}(z, u)]'$$

$$= A_{x,z} \Sigma_z - A_{x,z} A_{z,u} \Sigma_u A_{z,u}'$$

$$= A_{x,z} \left[\Sigma_z - A_{z,u} \Sigma_u A_{z,u}' \right]$$

$$= A_{x,z} \Sigma_{z|u}, \text{ using (4.5)}$$

$$(iii) \quad \text{Also, } A_{(z,x)|u} \Sigma_{x|u} = \text{Cov}(z, x | u), \text{ by definition}$$

$$= [\text{Cov}(x, z | u)]'$$

$$= \Sigma_{z|u} A_{x,z}, \text{ from result above}$$

$$\text{Hence, } A_{(z,x)|u} = \Sigma_{z|u} A_{x,z} \Sigma_{x|u}^{-1}$$

$$(iv) \quad A_{x,u} = A_{x,z} A_{z,u}, \text{ from theorem 4.1,}$$

$$= \left[\Sigma_{z|u}^{-1} A_{(z,x)|u} \Sigma_{x|u} \right] A_{z,u} \text{ using (4.7)}$$

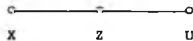
$$= \Sigma_{x|u} A_{(z,x)|u} \Sigma_{z|u}^{-1} A_{z,u} \quad \square$$

4.3 Comments and Diagrammatic Illustration

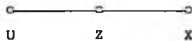
4.3.1 Conditional Independence Diagrams

(i) For the general case where X, Z, U are any three vectors such that $X \perp\!\!\!\perp U | Z$, the following diagrams (undirected

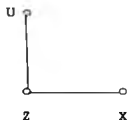
graphs) illustrate the different possibilities with time as the variable on the horizontal axis.



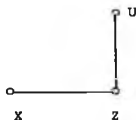
(a)



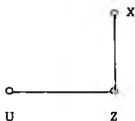
(b)



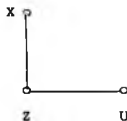
(c)



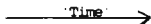
(d)



(e)

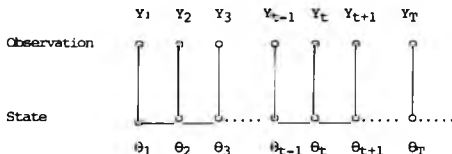


(f)



In diagrams (a) and (b), X , Z and U can all be states in a DLM, whilst in diagrams (c), (d), X , Z are states with U an observation vector. Similar comments apply to diagrams (e) and (f).

- (ii) We consider here the special case of the Dynamic Linear Models:



Here, we note that

- (a) $\theta_t \perp\!\!\!\perp \theta_{t+j} \mid \theta_{t+k}$, for all $0 < k < j$, and
- (b) $Y_t \perp\!\!\!\perp \theta_{t+i}, Y_{t+i} \mid \theta_t$, for all $i \neq 0$.

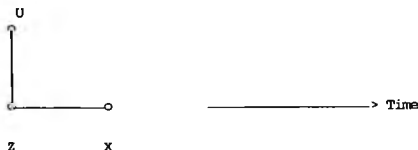
This conditional structure in which the future and the past are conditionally independent given the present confirm, of course, the Markovian property characteristic of the parameters/states of the DLM. More important, it helps to develop recurrence relationships for forecasting, retrospective analysis and the calculation of diagnostics.

4.3.2 Incorporating Information

Theorem 4.2 provides very powerful results which are of major practical significance. The expressions (4.4) and (4.5) for $\mu_{x|u}$ and $\Sigma_{x|u}$ in terms of $\mu_{z|u}$ and $\Sigma_{z|u}$ respectively describe the fact that there is passage of information from U to Z and then from Z to X . In more practical terms, the information about

X provided by U may be derived simply from posterior distribution of $Z|U$, together with the prior regression matrix of X on Z . Consequently, these results are of major importance in dynamic modelling and provide a general foundation for incorporating information whether it be observational, subjective, or obtained from other forecasting systems.

The results (4.6) and (4.7) of Theorem 4.2 are particularly revealing. Results (4.6) and (4.7) say that $A_{x,z} = A_{(z,x)}|u$, i.e. $A_{x,z}$ does not change given U , and generally $A_{z,x} \neq A_{(z,x)}|u$, i.e. $A_{z,x}$ does generally change given U . Consider the conditional independence diagram for the three vectors U , Z , X with $X \perp\!\!\!\perp U | Z$.



Let U represent new information, say the deletion of an observation. If we move forward in time, $A_{x,z} = A_{(x,z)}|u$ implies that the usual updating of the states via the regression coefficients following the deletion would not be affected, provided that $X \perp\!\!\!\perp U | Z$. In the case of DLMS, Z , X would be states.

However, if we move backward in time, $A_{z,x} \neq A_{(z,x)}|u$

implies that any filtering distribution would be affected by the deletion. These are intuitively very interesting results; they will be useful for the development of further results related to deletion of observations in Chapter 5.

The results of Theorem 4.1 and Corollary 4.1 are equally of major practical significance. Bearing in mind these comments and those made in subsection 4.3.1(i), we shall now consider some special cases to illustrate how information is incorporated using these theorems.

4.4 Routine Learning and Updating in the NDLM

Using results (4.4) and (4.5) of Theorem 4.2, we can now obtain in a fairly straightforward manner the updating equations for the state vector in the case of the Normal Dynamic linear model with known variance, i.e. $\phi = 1$.

We write $Z = U = Y_n$ and $X = \theta_n$, and everything is conditioned on D_{n-1} . Then $\theta_n \perp\!\!\!\perp Y_n \mid Y_n$.

Using (4.4), the updating equation for the mean of the state vector is given by

$$\mu_{x|u} = \mu_x + A_{x,z}(\mu_{z|u} - \mu_z),$$

with $\mu_{x|u} = a_{n,n} = E(\theta_n \mid D_n)$, the posterior mean

$$\mu_x = a_{n-1,n} = E(\theta_n \mid D_{n-1}), \text{ the prior mean}$$

$A_{x,z}$ = Regression matrix of θ_n on Y_n given $D_{n-1} = A_n$

$$\mu_z | u = E(Y_n | Y_n) = Y_n$$

$$\mu_z = E(Y_n | D_{n-1}) = \text{one-step ahead forecast mean}$$

Hence, we have the standard updating equation

$$a_{n,n} = a_{n-1,n} + A_n(Y_n - \mu_z) = a_{n-1,n} + A_n e$$

Similarly, using (4.5), the updating equation for the variance of the state vector is given by

$$\Sigma_x | u = \Sigma_x + A_{x,z}(\Sigma_z | u - \Sigma_z) A_{x,z}'$$

with $\Sigma_x | u = R_{n,n} = \text{Var}(\theta_n | D_n)$, the posterior variance

$$\Sigma_x = R_{n-1,n} = \text{Var}(\theta_n | D_{n-1}), \text{ the prior variance}$$

$$\Sigma_z | u = \text{Var}(Y_n | Y_n) = 0$$

$$\Sigma_z = \text{Var}(Y_n | D_{n-1}) = Q_n = \text{variance of one-step ahead forecast distribution.}$$

Hence we have the standard updating equation

$$R_{n,n} = R_{n-1,n} - A_n Q_n A_n'$$

4.5 Retrospective Analysis in the NDLM

From the theorems, the retrospective backward recurrence relations are easily obtained when, for all $t < n$, $X = \theta_t$, $Z = \theta_{t+1}$ and $U = Y_n$. Then obviously $\theta_t \perp\!\!\!\perp Y_n \mid \theta_{t+1}$, so that the results hold with known variance $\phi = 1$. Moreover, the following theorem provides a simple derivation of the full state distribution, whether or not the observation variances are known up to the scalar ϕ .

Theorem 4.3

For the Normal Dynamic Model as defined in Subsection 3.5, given D_n for all n and $t < t+k \leq n$,

- (i) the regression matrices $A_{t,t+k}$ of θ_t on θ_{t+k} remain constant and

$$A_{t,t+1} = R_{t,t} G_{t+1}^{-1} \left(G_{t+1} R_{t,t} G_{t+1}^{-1} + W_{t+1} \right)^{-1}, \quad (4.9)$$

- (ii) given ϕ , the joint distribution of the historic states is normal and is defined by the marginal distributions $\theta_t \mid D_n$, $\phi \sim N(a_{n,t}; R_{n,t}/\phi)$ and the covariances

$$\text{Cov}(\theta_t, \theta_{t+i} \mid D_n, \phi) = \left(\prod_{j=0}^{i-1} A_{t+j,t+j+1} \right) R_{n,t+i}/\phi, \quad i > 0 \quad (4.10)$$

where starting with $a_{n,n}$ and $R_{n,n}$, the moments may be calculated

using the backward recurrence relations:

$$a_{n,t} = a_{n-1,t} + A_{t,t+1}(a_{n,t+1} - a_{n-1,t+1}) , \quad (4.11)$$

$$R_{n,t} = R_{n-1,t} + A_{t,t+1}(R_{n,t+1} - R_{n-1,t+1}) A'_{t,t+1} \quad (4.12)$$

$$(iii) \quad v_{n-1} S_{n-1} \phi \mid D_{n-1} \sim \chi^2_{v_{n-1}} ,$$

$$v_n S_n \phi \mid D_n \sim \chi^2_{v_n} , \text{ where}$$

$$v_n = v_{n-1} + h , \quad (4.13)$$

$$v_n S_n = v_{n-1} S_{n-1} + e'_{n-1} Q_n^{-1} e_n , \quad (4.14)$$

and the marginal distributions of θ 's are multivariate T.

Proof:

$$(i) \quad A_{t,t+1} = \text{Cov}(\theta_t, \theta_{t+1}) [\text{Var}(\theta_{t+1})]^{-1} , \text{ by definition} \\ = \left[\text{Cov}(\theta_{t+1}, \theta_t) \right]' \left[G_{t+1} R_{t,t} G'_{t+1} + W_t \right]^{-1} ,$$

using the definition of the system equation from Subsection 2.4.1 of Chapter 2,

$$= \left[G_{t+1} \text{Cov}(\theta_t, \theta_t) \right]' \left[G_{t+1} R_{t,t} G'_{t+1} + W_t \right]^{-1} ,$$

again using the system equation

$$= R_{t,t} G'_{t+1} G_{t+1} R_{t,t} G_{t+1} + W_t^{-1}$$

From the result (4.6) of Theorem 4.2, it follows that the regression matrices $A_{t,t+k}$ of θ_t on θ_{t+k} remain constant.

(ii) From West and Harrison (1989a) (Chapter 4), it is established that, given ϕ , the joint distribution of the historical states is normal, and is defined by the marginal distributions

$$\theta_t | D_n, \phi \sim N(a_{n,t}; R_{n,t}/\phi).$$

Writing $X = \theta_t$, $Z = \theta_{t+1}$, $U = Y_n$, we have then $X \perp\!\!\!\perp U | Z$, so that theorem 4.2 applies and from (4.6),

$$\begin{aligned} \text{Cov}(\theta_t, \theta_{t+1} | D_n, \phi) &= \text{Cov}(\theta_t, \theta_{t+1} | D_{n-1}, Y_n, \phi) \\ &= A_{t,t+1} \text{Var}(\theta_{t+1} | D_n, \phi) \\ &= \left(\prod_{j=0}^{i-1} A_{t+j,t+j+1} \right) R_{n,t+1}/\phi, \text{ using} \\ &\quad (4.1). \end{aligned}$$

It is understood from the above and in what is to follow that everything is conditioned on D_{n-1} . Conditioning on Y_n implies conditioning on $\{D_{n-1}, Y_n\}$ i.e. D_n . To prove (4.11), we write $X = \theta_t$, $Z = \theta_{t+1}$, $U = Y_n$, so that again $X \perp\!\!\!\perp U | Z$. Then from

(4.4), we have

$$\mu_{\theta_t} | D_n = \mu_{\theta_t} | D_{n-1} + A_{t,t+1} (\mu_{\theta_{t+1}} | D_n - \mu_{\theta_{t+1}} | D_{n-1})$$

i.e.

$$a_{n,t} = a_{n-1,t} + A_{t,t+1} (a_{n,t+1} - a_{n-1,t+1})$$

And from (4.5), we have

$$\Sigma_{\theta_t} | D_n = \Sigma_{\theta_t} | D_{n-1} + A_{t,t+1} (\Sigma_{\theta_{t+1}} | D_n - \Sigma_{\theta_{t+1}} | D_{n-1}) A'_{t,t+1}$$

i.e.

$$R_{n,t} = R_{n-1,t} + A_{t,t+1} (R_{n,t+1} - R_{n-1,t+1}) A'_{t,t+1}$$

It is to be noted that, in the above expression, ϕ cancels out.

(iii) The proof is straightforward and follows from (3.18) - (3.20).

Thus, we have $v_{n-1} S_{n-1} \phi | D_{n-1} \sim \chi^2_{v_{n-1}}$, and

$$v_n S_n \phi | D_n \sim \chi^2_{v_n}, \text{ with } D_n = \{D_{n-1}, Y_n\},$$

where $v_t = v_{n-1} + h$,

$$v_n S_n = v_{n-1} S_{n-1} + e'_n Q_n^{-1} e_n, \text{ where } Q_n = \text{Var}(Y_n | D_{n-1}). \quad \square$$

CHAPTER FIVE

DELETION OF INFORMATION

5.1 Introduction

In Chapters 3 and 4, we study the combination of information with historic information in order to update forecasts, and to reassess what happened in the past. This routine operation of incorporating information holds as long as the model is found to be satisfactory according to an appropriate monitoring mechanism. There occur situations when the monitoring mechanism detects suspect data sets, which may then lead to a desire to eliminate a subset of information. For example, such a subset may relate to wars in economics or strikes in commerce and industry.

This chapter deals with the deletion of a single observation and then of a set of observations; appropriate recurrence relationships for the first two moments of the resulting distributions of the state vector are derived. And finally, a stochastic variance model is also considered together with a method of deriving its current jackknifed distribution.

5.2 Deletion of a Single Observation

5.2.1 Preliminary Comments and Notation

In line with the development of ideas in the preceding chapters, the regression matrix of the state vector on the set of

available information plays an important part in the derivation of appropriate recurrence relationships.

Consider the deletion of observation y_t in the information set in D_n ; write $D = D_n - y_t$. For $0 < t \leq \tau \leq n$, the regression matrix of θ_t on Y_τ given D is equal to:

Regression matrix of θ_t on θ_τ x regression matrix of θ_τ on Y_τ

$$= A_{t,\tau} A_\tau, \text{ since } \theta_t \perp\!\!\!\perp y_\tau \mid \theta_\tau \text{ and using (4.1)}$$

$$= A_{t,\tau} R_{n,\tau} F_\tau V_\tau^{-1}, \text{ using (3.8) from Lemma of Chapter 3.}$$

For $t \geq \tau$, $A_{t,\tau}$ changes as proved in Theorem 4.2 as per result (4.6). An alternative expression for $A_{t,\tau}$ is then derived in Theorem 5.1 in the next sub-section.

We now introduce some new notation which corresponds to (3.10a), (3.10b) and (3.10c), the ones used in subsection 3.4.1 when the deletion of information in normal dynamic models was first discussed. Write

$$d_t = y_t - F_t' a_{n,\tau}; \quad q_t = (V_t - F_t' R_{n,\tau} F_t)^{-1}; \quad B_t = R_{n,\tau} F_t' q_t^{-1} \quad (5.1)$$

Using the results (3.11) and (3.12) of theorem 3.1, we have

$$\theta_\tau \mid D, \emptyset \sim N(a_{n,\tau}^*, R_{n,\tau}^*/\emptyset) \text{ where}$$

$$a_{n,\tau}^* = a_{n,\tau} - B_\tau d_\tau \quad \text{and} \quad R_{n,\tau}^* = R_{n,\tau} + B_\tau F_\tau' R_{n,\tau} \quad (5.2)$$

5.2.2 Main Results

Since $A_{t,t} = I$ and, for $t < \tau \leq n$,
 $A_{t,\tau} = A_{t,t+1} A_{t+1,\tau}$, the full jackknifed state distribution may be
 obtained from the regression matrices $A_{t,t+1}$ and the marginal
 distributions calculated as in the following theorem.

Theorem 5.1

$\theta_t \mid D, \phi \sim N(a_{n,t}^*, R_{n,t}^* / \phi)$, $v \mid \phi \mid D \sim \chi_v^2$ and

$v_n s_n \mid D_n \sim \chi_{v_n}^2$, where

$$(i) \quad a_{n,t}^* = a_{n,t} + A_{t,\tau} (a_{n,\tau}^* - a_{n,\tau}), \quad (5.3)$$

$$R_{n,t}^* = R_{n,t} + A_{t,\tau} (R_{n,\tau}^* - R_{n,\tau}) A_{t,\tau}. \quad (5.4)$$

$$(ii) \quad \text{and for } t \geq \tau, \quad A_{t,\tau} = R_{n,t} A_{\tau,t} R_{n,\tau}^{-1}. \quad (5.5)$$

$$(iii) \quad v = v_n - h, \quad (5.6)$$

$$vs = v_n s_n - d_t' q_t^{-1} d_t. \quad (5.7)$$

Proof:

We note that everything is conditioned on D , except when
 stated otherwise.

(i) Whether $t \leq \tau$ or $t \geq \tau$, we have $\theta_t \mid Y_t \mid \theta_\tau$, so
 that results (4.4) and (4.5) of Theorem 4.2 hold. With $x \mid U \mid z$,

we have from (4.4)

$$\mu_x | u = \mu_x + A_{x,z}(\mu_z | u - \mu_z)$$

$$\text{i.e.} \quad E(\theta_t | D_n) = E(\theta_t | D) + A_{t,\tau} \left[E(\theta_\tau | D_n) - E(\theta_\tau | D) \right],$$

$$\text{i.e.} \quad a_{n,t} = a_{n,t}^* + A_{t,\tau}(a_{n,\tau} - a_{n,\tau}^*)$$

$$\text{i.e.} \quad a_{n,t}^* = a_{n,t} + A_{t,\tau}(a_{n,\tau}^* - a_{n,\tau})$$

From (4.5), we have

$$\Sigma_x | u = \Sigma_x + A_{x,z} \left(\Sigma_z | u - \Sigma_z \right) A_{x,z}$$

$$\text{i.e.} \quad \text{Var}(\theta_t | D_n) = \text{Var}(\theta_t | D) + A_{t,\tau} \text{Var}(\theta_\tau | D_n) - \text{Var}(\theta_\tau | D) A_{t,\tau}$$

$$\text{i.e.} \quad R_{n,t} = R_{n,t}^* + A_{t,\tau}(R_{n,\tau} - R_{n,\tau}^*) A_{t,\tau}$$

$$\text{i.e.} \quad R_{n,t}^* = R_{n,t} - A_{t,\tau}(R_{n,\tau} - R_{n,\tau}^*) A_{t,\tau}$$

$$= R_{n,t} + A_{t,\tau}(R_{n,\tau}^* - R_{n,\tau}) A_{t,\tau}$$

(ii) For the case $t \geq \tau$, the regression matrices vary with n , as proved in result (4.7) of Theorem 4.2 and as discussed in subsection 4.3 of Chapter 4. So we use result (4.7) here to express $A_{t,\tau}$ in terms of constant regression matrices $A_{t,t}$.

To be in line with result (4.7), write $\theta_t = X$, $Y_\tau = U$,

$\theta_{\tau} = z$, so that $x \perp\!\!\!\perp u \mid z$ and

$$\Sigma_{x|u} = \Sigma_{\theta_{\tau}|D_n} = \text{Var}(\theta_{\tau} | D_n) = R_{n,\tau}, \text{ since } D_n = \{D, Y_{\tau}\}$$

$$\Sigma_{z|u}^{-1} = R_{n,\tau}^{-1},$$

$$A'_{(z,x)|u} = A'_{(\tau,t) | D_n} = A'_{(\tau,t) | D}, \text{ from result (4.6)}$$

so that $A_{\tau,\tau} = R_{n,\tau} A'_{\tau,\tau} R_{n,\tau}^{-1}$

(iii) The result (3.21) of Theorem 3.2 is relevant to prove the result (5.7) here. The setting is similar and the additive property of the χ^2 -distribution makes it possible to use the well-known results obtained in learning on \emptyset .

But y_{τ} is not the last observation as u was then. Moreover, given D_n , the probability distributions are independent of the order in which the observations have been processed, by virtue of symmetry. So, considering y_{τ} to be processed last, then result (3.21) applies, giving

$$v = v_n - h \text{ and}$$

$$v_s = v_n s_n - d'_{\tau} g_{\tau}^{-1} d_{\tau}. \quad \square$$

Corollary 5.1

The results (5.3) and (5.4) may be expressed sequentially, dualing the updating recurrences (4.11) and (4.12), so that for $t \leq \tau$,

$$a_{n,t}^* = a_{n,t} + A_{t,t+1}(a_{n,t+1}^* - a_{n,t+1})$$

$$R_{n,t}^* = R_{n,t} + A_{t,t+1}(R_{n,t+1}^* - R_{n,t+1}) A'_{t,t+1}$$

Proof:

The result is obvious by taking $\tau = t + 1$ in (5.3) and (5.4). □

5.3 Finite Truncation Model

Sometimes practitioners like to discard all information which is older than a given age; for example, such information may be considered undesirable in terms of their effect on any forecast. Then the resulting finite truncation model of length l observations bases all forecasts on $D_n^* = \{Y_{n-l+1}, \dots, Y_n\}$.

Now for information set $D_{n-1}^* = \{Y_{n-l}, Y_{n-l+1}, \dots, Y_{n-1}\}$, we have $\theta_t | D_{n-1}^*, \phi \sim N(a_{n-1,t}, R_{n-1,t} / \phi)$, with known moments. Observation Y_n is received, so that observation Y_{n-l} is to be deleted. We note that $D_n^* = \{D_{n-1}^* - Y_{n-l}, Y_n\}$, and also that $t > \tau = n - l$, so that result (5.5) is relevant. In a

straightforward way, we can apply results (5.3), (5.4) and (5.5) to perform the deletion of y_{n-1} from the information set D_{n-1}^* and then produce new moments $a_{n-1,t}^*$ and $R_{n-1,t}^*$. Then, in turn, upon receipt of y_n , the usual updating of moments is carried out to obtain $a_{n,t}$ and $R_{n,t}$.

Alternatively, routine updating of $a_{n-1,t}$ and $R_{n-1,t}$ can be performed upon receipt of y_n and then the deletion of y_{n-1} from information set D_n^* carried out using results (5.3), (5.4) and (5.5).

The procedure is then repeated as observations y_{n+i} are received for $i > 0$.

5.4 Deletion of a set of l Observations

5.4.1 Introduction

Whenever a particular subset of observations is found to be suspect especially following model breakdown as detected by a given model monitoring mechanism, there is a need to delete that given set and compute the moments of the resulting distributions of the state vector. The recurrence relationships (5.3) - (5.7) can be used to that effect in a sequential manner.

5.4.2 Methodology/Procedure

Consider the case when a set, $\{y_{\tau_1}, \dots, y_{\tau_l}\}$, of l observations, where $\tau_i < \tau_{i+1}$, is to be deleted. Write

$D_{n,i} = D_n - \{y_{\tau_1}, \dots, y_{\tau_i}\}$. For example suppose the current distribution $\theta_n | D_{n,l}, \emptyset$ and $\theta_n, \emptyset | D_{n,l}$ are desired. This is achieved quickly using a sequential procedure which requires calculations only at the $l+1$ time points $\tau_1, \tau_2, \dots, \tau_l, \tau_{l+1}$. To obtain maximum clarity, the procedure is presented on a step-by-step basis.

(i) Delete y_{τ_1} .

Then use (5.1) to derive d_{τ_1}, q_{τ_1} , defined by

$$d_{\tau_1} = y_{\tau_1} - F'_{\tau_1} a_{n,\tau_1}, \text{ and}$$

$$q_{\tau_1} = V_{\tau_1} - F'_{\tau_1} R_{n,\tau_1} F_{\tau_1}$$

Thereafter we obtain $(\theta_{\tau_1} | D_{n,1}, \emptyset) \sim N(a_{n,\tau_1}^*, R_{n,\tau_1}^* / \emptyset)$ where

$a_{n,\tau_1}^*, R_{n,\tau_1}^*$ are defined as per (5.2). Using (5.3), (5.4),

(5.5), $(\theta_{\tau_2} | D_{n,1}, \emptyset)$ is obtained with the two moments defined,

since $\theta_{\tau_2} \perp\!\!\!\perp y_{\tau_1} | \theta_{\tau_1}$.

(ii) Additionally delete y_{τ_2} .

Use (5.1) to derive the corresponding $d_{\tau_2}^*, q_{\tau_2}^*$ and

use (5.2) to obtain

$$(\theta_{\tau_2} | D_{n,2}, \emptyset) \sim N(a_{n,\tau_2}^*, R_{n,\tau_2}^* / \emptyset)$$

Again the conditional independence property $\theta_{\tau_3} \perp\!\!\!\perp y_{\tau_2} | \theta_{\tau_2}$ ensures that, using theorem 5.1 and results (5.3), (5.4) and (5.5), $\theta_{\tau_3} | D_{n,2}, \emptyset$

is obtained with the first two moments defined.

(iii) Additionally delete y_{τ_3} .

Derive $d_{\tau_3}^*$, $q_{\tau_3}^*$ using (5.1) and the procedure is repeated. The conditional independence property $\theta_{\tau_{i+1}} \perp\!\!\!\perp y_{\tau_i} \mid \theta_{\tau_i}$ ensures that the procedure can continue.

(iv) For the final application of the procedure, we use (5.3), (5.4) and (5.5) to give $(\theta_n \mid D_{n_{\tau_l}}, \emptyset)$ from $(\theta_{\tau_l} \mid D_{n_{\tau_l-1}}, \emptyset)$ (noting $n = \tau_{l+1}$).

5.4.3 The Case of Unknown Constant Variance

There are two cases of unknown variance that we shall consider in the present context. The usual case, considered previously, has been when variance is considered unknown but constant; and this will be considered in this subsection. The next case, to be considered in section 5.5, deals with stochastic variation in variance.

The case of an unknown constant variance is straightforward if we make use of results (5.6) and (5.7) of Theorem 5.1. To start with, we have

$$v_n s_n \emptyset \mid D_n \sim \chi_{v_n}^2.$$

Then, performing the deletion of l observations using the procedure

described in Section 5.4.2, we have

$$v_{n,l} s_{n,l} \phi \mid D_{n,l} \sim \chi^2_{v_{n,l}},$$

where $v_{n,l}$ and $s_{n,l}$ are defined such that they correspond to data $D_{n,l}^*$ as defined in Section 5.4.2. Also, we have

$$v_{n,l} = v_n - lh, \quad (5.8)$$

$$v_{n,l} s_{n,l} = v_n s_n - \sum_{i=1}^l d_{t_i}^{*'} q_{t_i}^{*-1} d_{t_i}^*, \quad (5.9)$$

with (d_{t_1}, q_{t_1}) replaced by $(d_{t_1}^*, q_{t_1}^*)$.

5.5 The Case of Stochastic Variance

In the spirit of dynamic modelling discussed in Chapter 2, Section 2, the unknown scalar ϕ is often considered to change slowly through time; hence it is indexed by t , written ϕ_t . The model for such stochastic changes in variation developed in Ameen and Harrison (1985) and Harrison and West (1986) and presented in West and Harrison (1989a, Chapter 10, Section 10.8) can accommodate quite easily the procedure for deleting a set of observations.

First, the model is introduced briefly. On top of the usual model definition of the Dynamic Linear Model as presented in

Chapter 2, subsection 2.3, the stochastic variation in ϕ is modelled by a random walk,

$$\phi_t = \phi_{t-1} + \psi_t, \text{ with } \psi_t \sim [0, U_t].$$

The variance of the precision parameter, ϕ_t , is found to increase and this increase is described in a multiplicative way by making use of the discounting principle. Just as in the case of the modelling of the evolution of θ_t , the stochastic variation here is modelled by means of discount factors. These determine effectively the values of U_t .

The discount variance model operates as follows with a variance discount factor δ , usually such that $0.98 < \delta < 1$; the range of values for δ indicates that only small degrees of stochastic variation are being taken into account. The marginal distributions for ϕ_t may be modelled by a power steady model (Smith, 1979) applied to $\psi_t = \log \phi_t$ in which

$$p(\psi_t = \psi | D_{t-1}) \propto \{p(\psi_{t-1} = \psi | D_{t-1})\}^\delta.$$

With prior for ϕ defined by

$$v_{n-1} s_{n-1} \phi_{n-1} | D_{n-1} \sim \chi_{v_{n-1}}^2,$$

we note that

$$v_n s_n \phi_n \mid D_{n-1} \sim \chi^2_{\delta v_{n-1}},$$

so that the number of degrees of freedom effectively decreases; then we have the posterior defined by

$$v_n s_n \phi_n \mid D_n \sim \chi^2_{v_n},$$

where $v_n = \delta v_{n-1} + h$

and $v_n s_n = \delta v_{n-1} s_{n-1} + e_n' Q_n^{-1} e_n$.

Lemma 5.1

On deleting k observations as in Section 5.4, we have

$$v_{n,k} s_{n,k} \phi_n \mid D_{n,k} \sim \chi^2_{v_{n,k}}, \text{ where}$$

$$v_{n,k} = v_n - h \sum_{i=1}^k \delta^{n-\tau_i}, \quad (5.10)$$

$$v_{n,k} s_{n,k} = v_n s_n - \sum_{i=1}^k \delta^{n-\tau_i} d_{\tau_i}^{*'} q_{\tau_i}^{*-1} d_{\tau_i}^* \quad (5.11)$$

Proof:

The proof follows from (5.8) and (5.9) and the equivalence of the calculations for $\theta_n, \phi_n | D_{n,l}$ with those from the model

$$\{F_t, G_t, v_t/\theta^{n-t}, w_t/\theta^{n-t}; R_{0,1}/\theta^n\}, \text{ in which}$$

all the known variances relating to time $0 \leq t \leq n$ are replaced by those in the original model divided by θ^{n-t} . Some clarification is given below.

Consider, first, the deletion of one observation vector y_1 . Then from results of (5.6) and (5.7) in the case of deletion of y_1 when unknown constant variance is considered, we have now

$$v_{n,1} = v_n - h\theta^{n-1},$$

$$v_{n,1} s_{n,1} = v_n s_n - \theta^{n-1} d_1' q_1^{-1} d_1.$$

This says that for each observation, in the observation vector of length h , there is a loss of θ^{n-1} in the number of degrees of freedom each time the observation y_1 is deleted. The loss occurs correspondingly in the expression for $v_{n,1} s_{n,1}$.

Then, with l deletions, y_{1_1}, \dots, y_{1_l} , the loss in the

degree of freedom cumulates, so that we have

$$v_{n,l} = v_n - h \sum_{i=1}^l \delta^{n-\tau_i},$$

$$v_{n,l} s_{n,l} = v_n s_n - \sum_{i=1}^l \delta^{n-\tau_i} d_{\tau_i}^* q_{\tau_i}^{*-1} d_{\tau_i}^* . \quad \square$$

CHAPTER SIX

THE SPECIAL CASE OF DISCOUNT WEIGHTED REGRESSION AND APPLICATIONS

6.1 Introduction

This Chapter deals with the special case of Discount Weighted Regression (DWR) which is the generalisation of Exponentially Weighted Regression (EWR) as developed by Brown (1962). The ideas regarding DWR were first developed by Ameen and Harrison (1984, 1985). It is established, in the next section, that DWR is in fact equivalent to a class of DLMS. This makes it possible to apply the results for incorporation and deletion of information, developed so far, to DWR and EWR. There is the added advantage that DWR dynamic models offer very simple and neat procedures. Furthermore these models provide a link with static models and least squares procedures.

In the next section, after defining DWR, we establish the equivalence between DWR and the DLM. In the following two sections, recurrence relationships for the incorporation and deletion of information in certain special cases are then derived from results obtained in Chapters 3, 4 and 5. We end up in the last section with an illustration of the results using a data set.

6.2 Discount Weighted Regression6.2.1 Definition

In light of what has been discussed above, it is appropriate to define first Exponentially Weighted Regression, in the simple univariate case. EWR is characterised by a local model defined by

$$Y_n = F_n' \theta + v_n, \quad n = 0, 1, \dots \quad (6.1)$$

where, as usual, Y_n , F_n , θ are the observations, the vector of independent variables, the vector of parameters respectively and v_n , error terms, identically and independently distributed with $v_n \sim N(0, V_n)$. Then given a discount factor δ , with $0 < \delta \leq 1$, the parameter vector θ is estimated by $\hat{\theta}_n$ where $\hat{\theta}_n$ minimises the discounted sum of squares

$$S_n(\theta) = \sum_{i=0}^{n-1} \delta^i (y_{n-i} - F_{n-i}' \theta)^2 \quad (6.2)$$

Thus, the latest observation, y_n , is given the maximum weight of 1 and the first observation is given the least weight δ^{n-1} which would obviously tend to zero for large n .

It is to be noted that

$$S_n(\theta) = (y_n - F_n' \theta)^2 + \delta S_{n-1}(\theta), \quad (6.3)$$

since

$$\begin{aligned}
 S_n(\theta) &= (y_n - E'_n \theta)^2 + \sum_{i=1}^{n-1} \delta^i (y_{n-i} - E'_{n-i} \theta)^2 \\
 &= (y_n - E'_n \theta)^2 + \delta \sum_{i=0}^{n-2} \delta^i (y_{n-1-i} - E'_{n-1-i} \theta)^2 \\
 &= (y_n - E'_n \theta)^2 + \delta S_{n-1}(\theta)
 \end{aligned}$$

Discount Weighted Regression is a generalisation of EWR whereby a general weight, v_n^{-1} , is used on top of the discounting principle. The discount factor, δ_n , is now no longer the n^{th} power of a constant δ ; it can vary for all values of n . Formally, the estimate $\hat{\theta}_n$ of θ is obtained by minimizing

$$S_n(\theta) = \sum_{i=0}^{n-1} v_{n-i}^{-1} \Delta_{n,i} (y_{n-i} - E'_{n-i} \theta)^2,$$

where $\Delta_{t,i} = \prod_{j=0}^{i-1} \delta_{t-j}$ for $i \geq 1$ and $\Delta_{t,0} = 1$, $\delta_n > 0$ and $v_n > 0$.

Following the above discussion, it is similarly established that

$$S_n(\theta) = v_n^{-1} (y_n - E'_n \theta)^2 + \delta_n S_{n-1}(\theta)$$

It is then evident that EWR and ordinary least squares regression are special cases of DWR. If $v_n = 1$ and $\Delta_{n,1} = \delta^1$,

with $\delta_{t-j} = \delta$, then the EWR case results; and if $V_n = 1$ and $\Delta_{n,i} = 1$, with $\delta_{t-j} = 1$, we have ordinary linear regression.

We can obtain the equivalent expressions in the case when Y_n is vector-valued. Then,

$$S_n(\theta) = \sum_{i=0}^{n-1} \Delta_{n-i} (Y_{n-i} - F_{n-i}' \theta)' V_{n-i}^{-1} (Y_{n-i} - F_{n-i}' \theta) \quad (6.4)$$

$$= (Y_n - F_n' \theta)' V_n^{-1} (Y_n - F_n' \theta) + \delta_n S_{n-1}(\theta), \quad (6.5)$$

where V_{n-i}^{-1} are $n \times n$ variance matrices,
 F_{n-i} are $n \times n$ matrices of known independent variables,
 Δ_{n-i} is as defined above,
 Y_n is a $n \times 1$ vector of observations,
 θ is a $n \times 1$ vector of states,
 δ_i 's is a set of known discount factors with
 $i = 1, 2, \dots, n$, with $0 < \delta_i \leq 1$.

6.2.2 Equivalence of DWR with the DLM

6.2.2.1 Definition and Notation

Let $S_n(\theta)$ be as defined in (6.4) and (6.5), together with the defined symbols as given in subsection 6.2.1. Let $D_n = \{y_1, y_2, \dots, y_n\}$. Then the point estimate, m_n , for θ given D_n is that value of θ which minimizes $S_n(\theta)$.

Define

$$(i) \quad X_n = F_n V_n^{-1} F_n' + \delta_n X_{n-1}, \quad \text{with } X_0 = 0$$

$$(ii) \quad H_n = F_n V_n^{-1} Y_n + \delta_n H_{n-1}, \quad \text{with } H_0 = 0$$

Also, let X_n be a full rank matrix and let

$$C_n = X_n^{-1}; \quad R_n = C_{n-1} / \delta_n;$$

$$A_n = R_n F_n Q_n^{-1}; \quad Q_n = V_n + F_n' R_n F_n$$

$$f_n = F_n' m_{n-1}; \quad e_n = Y_n - f_n.$$

The following results from multivariate joint distributions, (West and Harrison, 1989a, Chapter 16) with the usual notation and with subscripts deleted, will prove useful in the proof to be given:

$$C^{-1} = R^{-1} + FV^{-1}F' \quad (6.6)$$

$$C = R - RF[F'RF + V]^{-1}F'R \quad (6.7)$$

It is to be noted that (6.6) was given using different symbols as result (3.6) in Chapter 3 and was then used to prove Theorem 3.1. Also, C , R , V being variance matrices are full rank positive definite matrices.

6.2.2.2 Theorem 6.1

The value m_n of θ which minimises

$$S_n(\theta) = (y_n - F_n' \theta)' V_n^{-1} (y_n - F_n' \theta) + \delta_n S_{n-1}(\theta)$$

satisfies

$$(i) \quad m_n = C_n H_n ,$$

$$(ii) \quad m_n = m_{n-1} + \lambda_n e_n ,$$

$$(iii) \quad C_n = R_n - \lambda_n Q_n^{-1} \lambda_n .$$

Proof:

Using standard results of matrix differentiation, we have

$$\begin{aligned} \frac{\partial S_n(\theta)}{\partial \theta} &= -2 F_n' V_n^{-1} (y_n - F_n' \theta) + \delta_n \frac{\partial S_{n-1}(\theta)}{\partial \theta} \\ &= -2 [F_n' V_n^{-1} y_n - F_n' V_n^{-1} F_n' \theta] + \delta_n \frac{\partial S_{n-1}(\theta)}{\partial \theta} \\ &= -2 [H_n - \delta_n H_{n-1} - (X_n - \delta_n X_{n-1}) \theta] + \delta_n \frac{\partial S_{n-1}(\theta)}{\partial \theta} \\ &= -2 [H_n - \delta_n H_{n-1} - X_n \theta + \delta_n X_{n-1} \theta] + \delta_n \frac{\partial S_{n-1}(\theta)}{\partial \theta} \\ &= -2 [H_n - X_n \theta] + 2 \delta_n (H_{n-1} - X_{n-1} \theta) + \delta_n \frac{\partial S_{n-1}(\theta)}{\partial \theta} \end{aligned}$$

It is obvious that a recursive relationship is obtained.

$$\text{Further, } \delta_n \frac{\partial s_{n-1}(\theta)}{\partial \theta} = -2\delta_n(H_{n-1} - X_{n-1}\theta) + 2\delta_n\delta_{n-1}(H_{n-2} - X_{n-1}\theta) \\ + \delta_n\delta_{n-1} \frac{\partial s_{n-2}(\theta)}{\partial \theta}.$$

$$\text{Also, } \left(\prod_{j=0}^i \delta_{n-j} \right) \frac{\partial s_{n-i}(\theta)}{\partial \theta} = -2 \left(\prod_{j=0}^i \delta_{n-j} \right) (H_{n-i} - X_{n-i}\theta) \\ + 2 \prod_{j=0}^{i+1} \delta_{n-j} (H_{n-i-1} - X_{n-i-1}\theta) \\ + \prod_{j=0}^{i+1} \delta_{n-j} \frac{\partial s_{n-i-1}(\theta)}{\partial \theta}.$$

The last expression would be given by

$$\frac{\partial s_1(\theta)}{\partial \theta} = -2 \left(\prod_{j=0}^{n-1} \delta_{n-j} \right) (H_1 - X_1\theta) + 2 \left(\prod_{j=0}^n \delta_{n-j} \right) (H_0 - X_0\theta) \\ + \left(\prod_{j=0}^n \delta_{n-j} \right) + \frac{\partial s_0(\theta)}{\partial \theta} \\ = -2 \left(\prod_{j=0}^{n-1} \delta_{n-j} \right) (H_1 - X_1\theta),$$

by definition of X_0 , H_0 and S_0 .

Hence we have

$$\frac{\partial s_n(\theta)}{\partial \theta} = -2(H_n - X_n\theta),$$

so that, in turn, $\frac{\partial^2 S_n(\theta)}{\partial \theta^2} = 2X_n$.

Hence $\left| \frac{\partial^2 S_n(\theta)}{\partial \theta^2} \right| > 0$, by definition of X_n .

Thus a minimum for $S_n(\theta)$ is obtained when $\frac{\partial S_n(\theta)}{\partial \theta} = 0$,

i.e. when $X_n m_n = H_n$ (6.8)

i.e. when $m_n = X_n^{-1} H_n$,

i.e. when $m_n = C_n H_n$, thus proving (1) of theorem.

Now from (6.8), we have

$$\begin{aligned} 0 &= X_n m_n - H_n \\ &= X_n m_n - \delta_n H_{n-1} - F_n V_n^{-1} y_n, \text{ by definition of } H_n \end{aligned} \quad (6.9)$$

Also from (6.8),

$$\begin{aligned} 0 &= \delta_n (X_{n-1} m_{n-1} - H_{n-1}) \\ &= \delta_n X_{n-1} m_{n-1} - \delta_n H_{n-1} \\ &= (X_n - F_n V_n^{-1} F_n^T) m_{n-1} - \delta_n H_{n-1}, \text{ using definition of } X_n \\ &= X_n m_{n-1} - F_n V_n^{-1} F_n^T m_{n-1} - \delta_n H_{n-1} \end{aligned} \quad (6.10)$$

Then subtracting (6.10) from (6.9), we get

$$\begin{aligned} X_n (m_n - m_{n-1}) &= F_n V_n^{-1} (y_n - F_n^T m_{n-1}) \\ &= F_n V_n^{-1} e_n, \text{ by definition of } e_n \end{aligned}$$

Hence, $m_n = m_{n-1} + X_n^{-1} F_n' V_n^{-1} e_n$

$$= m_{n-1} + C_n F_n' V_n^{-1} e_n \quad (6.11)$$

But,

$$\begin{aligned} C_n F_n &= \{R_n - R_n F_n (F_n' R_n F_n + V_n)^{-1} F_n' R_n\} F_n, \text{ from (6.7)} \\ &= R_n F_n \{I - [F_n' R_n F_n + V_n]^{-1} F_n' R_n F_n\} \\ &= R_n F_n \{I - Q_n^{-1} F_n' R_n F_n\} \\ &= R_n F_n Q_n^{-1} (Q_n - F_n' R_n F_n) \\ &= A_n V_n, \text{ from definition of } A_n \text{ and } Q_n. \end{aligned}$$

Hence from (6.11), we have

$$m_n = m_{n-1} + A_n e_n, \text{ thus proving (2) of theorem.}$$

Finally from (6.7) we have

$$\begin{aligned} C_n &= R_n - R_n F_n (F_n' R_n F_n + V_n)^{-1} F_n' R_n \\ &= R_n - R_n F_n Q_n^{-1} F_n' R_n \\ &= R_n - R_n F_n Q_n^{-1} Q_n Q_n^{-1} F_n' R_n \\ &= R_n - A_n Q_n A_n', \text{ thus proving (3) of theorem.} \quad \square \end{aligned}$$

6.2.2.3 Conclusion

Noting that $R_n = C_{n-1}/\delta_n$, we find that the results of Theorem 6.1 are equivalent to the updating equations of the DLM $\{F_n, I, V_n, (\delta_n^{-1} - 1) C_{n-1}\}$, thus establishing the appropriate equivalence.

It is relevant to note that in the case of DLM, we have

$R_n = C_{n-1} + W_n$, so that, with $G_n = I$ we have

$$\begin{aligned} R_n &= C_{n-1} + (\delta_n^{-1} - 1) C_{n-1} \\ &= C_{n-1}[1 + \delta_n^{-1} - 1] \\ &= C_{n-1}/\delta_n. \end{aligned}$$

We shall now use the results of Chapters 4 and 5 dealing with the incorporation and deletion of information in dynamic models to establish simple, neat corresponding results for DWR and EWR.

6.3 Incorporation of Information

We shall consider, in particular, the retrospective analysis for Discount Weighted Regression and derive results which correspond to those of Theorem 5.1. We recall that DWR is equivalent to the DLM $\{F_t, I, V_t/\delta^{n-t}, Q\}$ (in the special case of static regression). Thus, $G_t = I$ for all t and $W_t = 0$. For such a model given D_n and for all n and $t < t + k \leq n$, we have

$$\begin{aligned}
 \text{(i)} \quad \text{Regression matrix of } \theta_t \text{ on } \theta_{t+1} &= A_{t,t+1} \\
 &= R_{t,t} G_{t+1}^{-1} R_{t,t+1} \\
 &= R_{t,t} G_{t+1}^{-1} (R_{t,t} / \delta_{t+1})^{-1} \\
 &= \delta_{t+1} I
 \end{aligned}$$

$$\text{Hence, } A_{t,t+k} = \left(\prod_{i=1}^k \delta_{t+i} \right) I.$$

In the specific case of EWR, we have

$$A_{t,t+1} = \delta I,$$

and

$$A_{t,t+k} = \delta^k I.$$

(ii) Given ϕ , the joint distribution of the historic states is normal and is defined by the marginal distributions

$\theta_t | D_n, \phi \sim N(a_{n,t}, R_{n,t}/\phi)$ and the covariances

$$\begin{aligned}
 \text{Cov}(\theta_t, \theta_{t+i} | D_n, \phi) &= A_{t,t+i} R_{n,t+i}/\phi \\
 &= \left(\prod_{j=1}^i \delta_{t+j} \right) R_{n,t+i}/\phi
 \end{aligned}$$

In the specific case of EWR, we have $\delta^i R_{n,t+i}/\phi$. Starting with $a_{n,n}$ and $R_{n,n}$, the moments may be calculated by the backward recurrence relations given by (4.11) and (4.12).

$$\begin{aligned}
a_{n,t} &= a_{n-1,t} + A_{t,t+1}(a_{n,t+1} - a_{n-1,t+1}) \\
&= a_{n-1,t} + \delta_{t+1} \left[a_{n-1,t+1} + A_{t+1,t+2}(a_{n,t+2} - a_{n-1,t+2}) \right. \\
&\quad \left. - a_{n-1,t+1} \right], \text{ expressing } a_{n,t+1} \text{ as a recurrence} \\
&\quad \text{relation} \\
&= a_{n-1,t} + \delta_{t+1} \delta_{t+1} (a_{n,t+2} - a_{n-1,t+2}) \\
&= a_{n-1,t} + \left[\prod_{j=t+1}^n \delta_j \right] (a_{n,n} - a_{n-1,n}) \\
&= a_{n-1,t} + A_n e_n \prod_{n=t-1}^n \delta_j, \text{ using updating equations} \\
&\quad \text{of NDLM as in subsection 4.4}
\end{aligned}$$

In the case of EWR, we have

$$\begin{aligned}
a_{n,t} &= a_{n-1,t} + A_n e_n \delta^{n-t} \\
R_{n,t} &= R_{n-1,t} + A_{t,t+1}(R_{n,t+1} - R_{n-1,t+1}) A_{t,t+1} \\
&= R_{n-1,t} + \delta_{t+1}^2 \left[R_{n-1,t+1} + A_{t+1,t+2}(R_{n,t+2} - R_{n-1,t+2}) A_{t+1,t+2} \right. \\
&\quad \left. - R_{n-1,t+1} \right], \text{ expressing } R_{n,t+1} \text{ as a recurrence} \\
&\quad \text{relation} \\
&= R_{n-1,t} + \delta_{t+1}^2 \left[A_{t+1,t+2}(R_{n,t+2} - R_{n-1,t+2}) A_{t+1,t+2} \right] \\
&= R_{n-1,t} + \delta_{t+1}^2 \delta_{t+2}^2 (R_{n,t+2} - R_{n-1,t+2}) \\
&= R_{n-1,t} + \left[\prod_{j=t+1}^n \delta_j^2 \right] (R_{n,n} - R_{n-1,n})
\end{aligned}$$

$$= R_{n-1,t} - A_n Q_n A_n' \prod_{j=t+1}^n \delta_j^2, \text{ using updating equation} \\ \text{of NDLM as in subsection 4.4.}$$

In the case of EWR, we have

$$R_{n,t} = R_{n-1,t} - A_n' Q_n A_n' \delta_n^2(n-t).$$

6.4 Deletion of Information

6.4.1 Deletion of One Observation y_t

We derive the results which correspond to those of Theorem 5.1.

We recall that $D = D_n - y_t$ and that $\theta_t | D, \emptyset \sim N(a_{n,t}^*, R_{n,t}^* / \emptyset)$.

Then we have the following results:

$$(i) \quad a_{n,t}^* = a_{n,t} + A_{t,\tau}(a_{n,\tau}^* - a_{n,\tau}) \\ = a_{n,t} + \left[\prod_{j=t+1}^{\tau} \delta_j \right] [a_{n,\tau} + A_{\tau,\tau}(a_{n,\tau}^* - a_{n,\tau})] \\ = a_{n,t} + a_{n,\tau}^* \prod_{j=t+1}^{\tau} \delta_j, \text{ since } A_{\tau,\tau} = I.$$

$$(ii) \quad R_{n,t}^* = R_{n,t} + A_{t,\tau}(R_{n,\tau}^* - R_{n,\tau}) A_{t,\tau}' \\ = R_{n,t} + \prod_{j=t+1}^{\tau} \delta_j^2 R_{n,\tau} + A_{t,\tau}(R_{n,\tau}^* - R_{n,\tau}) A_{t,\tau}' \\ = R_{n,t} + R_{n,\tau}^* \prod_{j=t+1}^{\tau} \delta_j^2$$

$$\begin{aligned}
 \text{(iii)} \quad \text{for } t \geq \tau, \quad A_{t,\tau} &= R_{n,t} A_{\tau,t}^{-1} R_{n,\tau}^{-1} \\
 &= R_{n,t} R_{n,\tau}^{-1} \prod_{j=\tau+1}^t \delta_j
 \end{aligned}$$

In the case of EWR, the corresponding results are

$$\text{(i)} \quad a_{n,t}^* = a_{n,t} + a_{n,\tau}^* \delta^{\tau-t}$$

$$\text{(ii)} \quad R_{n,t}^* = R_{n,t} + R_{n,\tau}^* \delta^{2(\tau-t)}$$

$$\text{(iii)} \quad \text{for } t \geq \tau, \quad A_{t,\tau} = R_{n,t} R_{n,\tau}^{-1} \delta^{t-\tau}$$

The sequential results of corollary 5.1 follow in a very straightforward manner both for DWR and EWR.

6.4.2 Deletion of a Set of Observations

From the expression obtained in 6.4.1, the expressions for the case of the deletion of a set of observations follow immediately since they are obtained in a sequential manner as described in subsection 5.4.

6.5 Applications

The set of data considered here are from the National Health Service (NHS) and they represent the number of prescriptions over a period of twenty-one months. The model for the data is the univariate steady model as defined in West and Harrison (1989a), Chapter 2 and the data set is referred to as NHSTEADY.

The following gives a summary of the graphs together with some brief comments, where necessary.

1. Figure 6.1 : This graph of raw data reveals two change-points at time $t = 10$ (i.e. month 10, year 1) and time $t = 17$ (i.e. month 5, year 2). There is an obvious need for intervention in both cases. In the first case, improvements in the treatment have brought a permanent, substantial decrease in the number of prescriptions; in the second case, an epidemic brought a sudden, sharp increase, which was not sustained as expected. This second one is a clear case of an outlier.
2. Figures 6.2 and 6.3 : These show respectively the graph with intervention at the first point and no intervention at the second, and then with intervention at both points. The intervention at the first point brings a lower level, whilst the second intervention, of major interest to us, constitutes a straightforward deletion of the outlier. The difference between the two graphs shows clearly the influential nature of the outlier.

3. Thereafter, for all graphs to follow, the two interventions are included and further deletions are carried out : point at time $t = 3$ is first deleted, then additionally point at time $t = 6$ and, further, additionally point at time $t = 8$ is finally deleted. These points, just like others, not being of major interest, the impact of their deletion is not very marked. Nevertheless, the graphs do indicate this marginal impact.
4. The content of the graphs is briefly indicated here, their interpretation being obvious:

Figure 6.4 : Filtered mean and raw data.

Figure 6.5 : Filtered mean for raw data and for the case of first deletion.

Figure 6.6 : Filtered mean for raw data and for the case of first two deletions.

Figure 6.7 : Filtered mean for the cases of one deletion and of first two deletions.

Figure 6.8 : Filtered mean for the cases of first two deletions and of three deletions.

Figure 6.9 : Filtered mean for raw data and the case of three deletions.

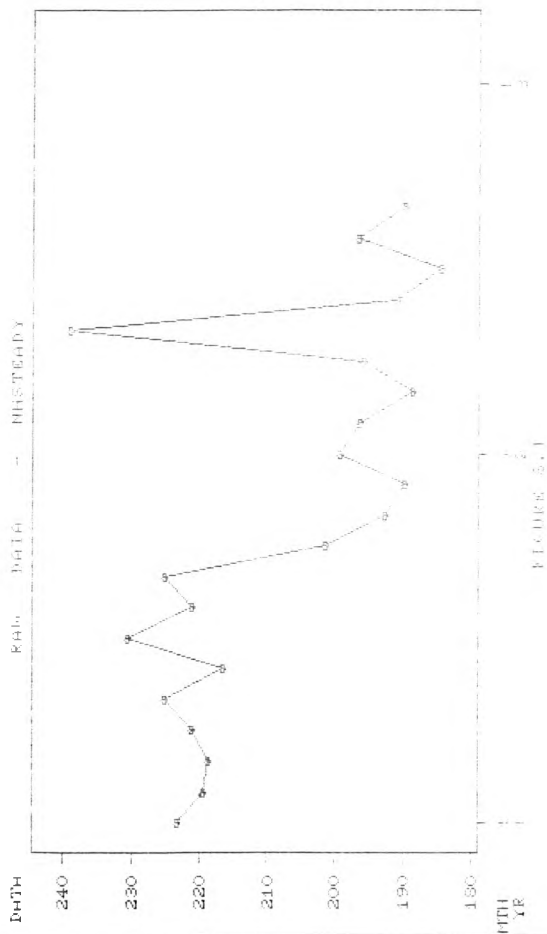
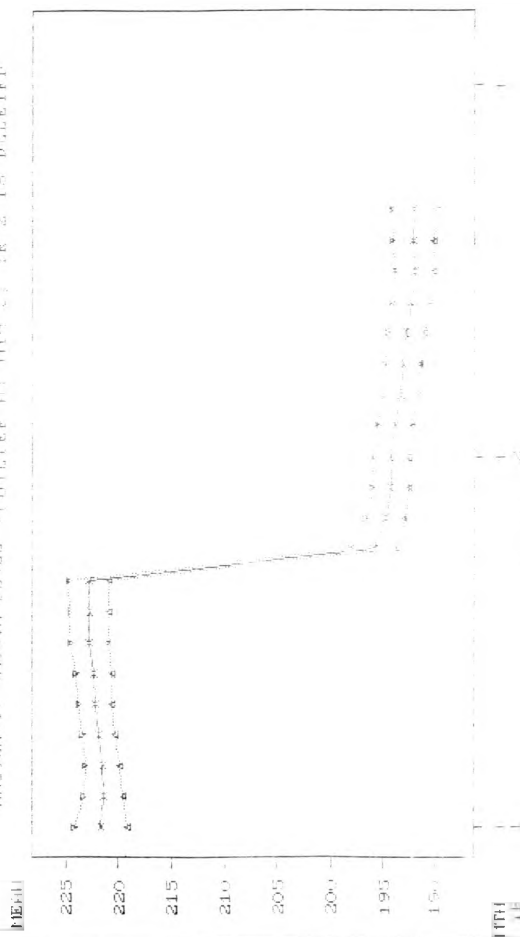


FIGURE 5.1

NOTE: POINT 1802 - OUTLIER OF 1100.5 IS DELETED



HISTOGRAM: APPROXIMATE COPIES OF THE 1000 OR 2 IS DELETED

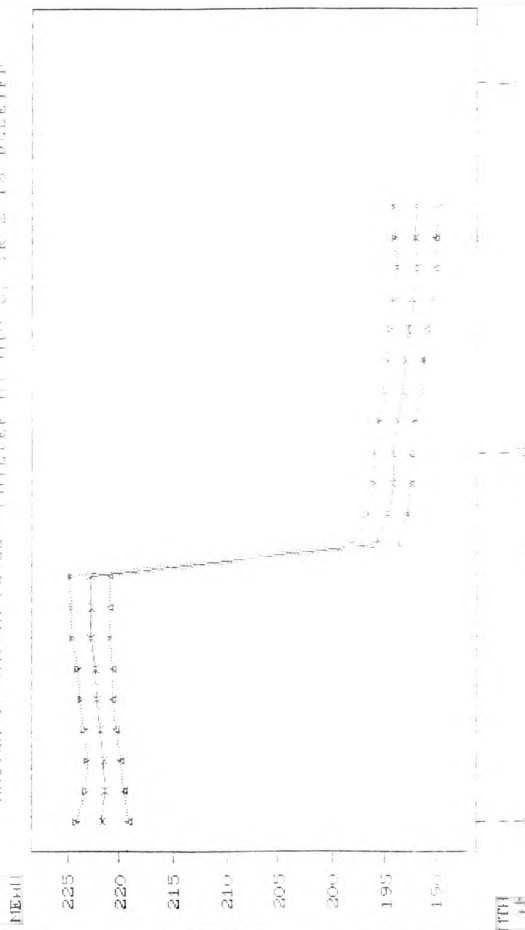


FIGURE 5.2

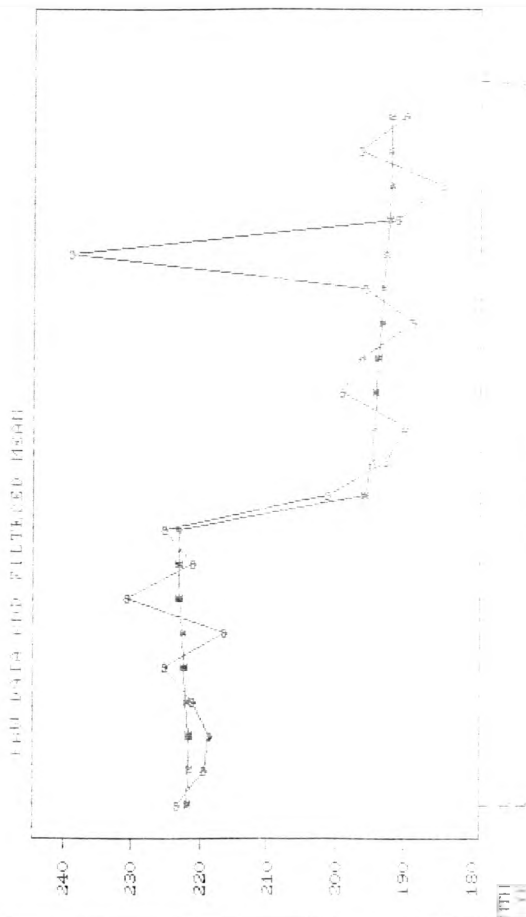


FIGURE 5.5

FILTERED HEAD FOR F&M DATA AND FOR ONE DELETION

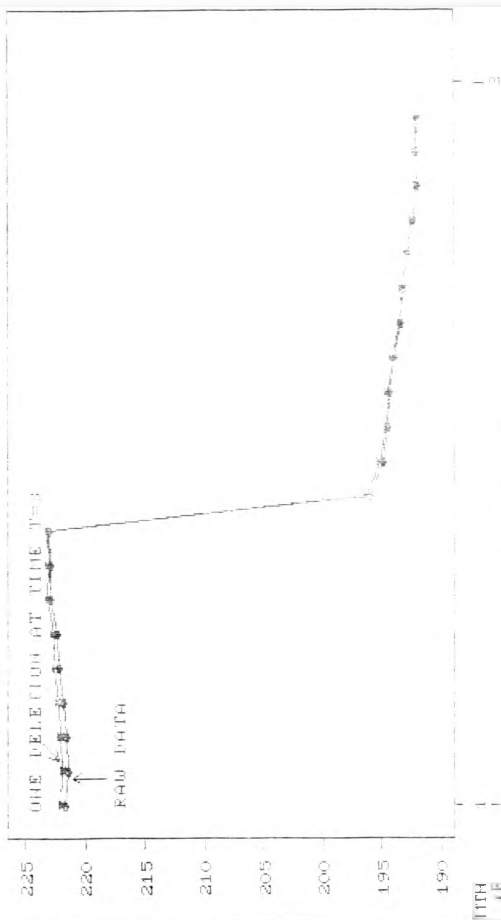


FIGURE 3.2

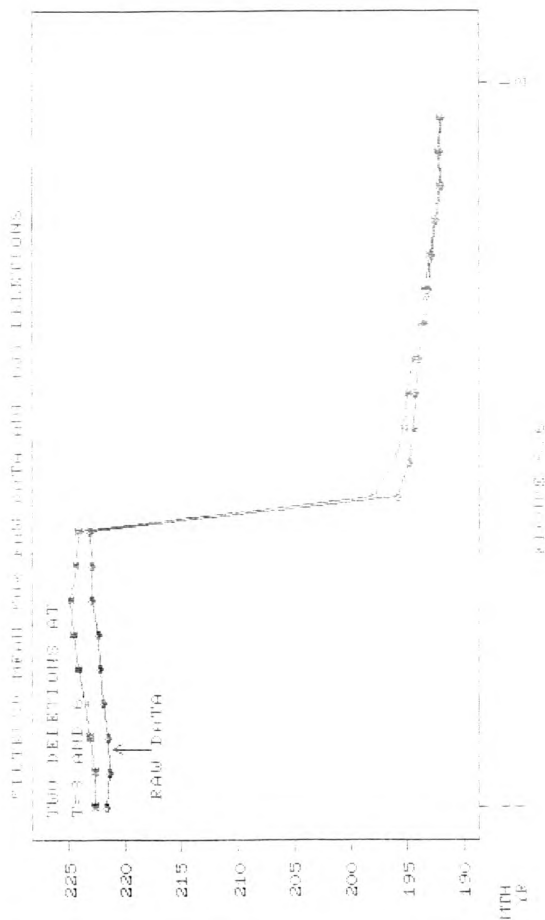


FIGURE BEHE FOR ONE DELETION AND TWO DELETIONS

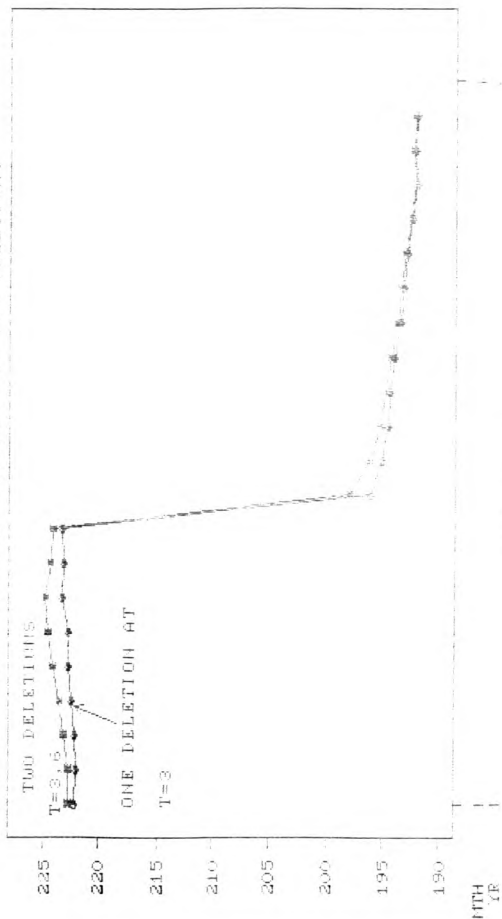
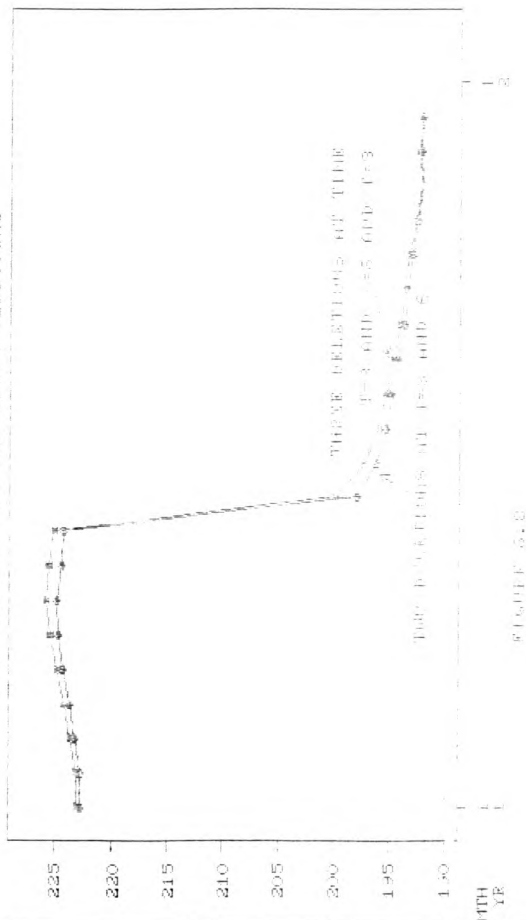


FIGURE 5.7

THE CASES OF TWO AND THREE DELETIONS



CASES OF NO DELETION AND OF THREE DELETIONS

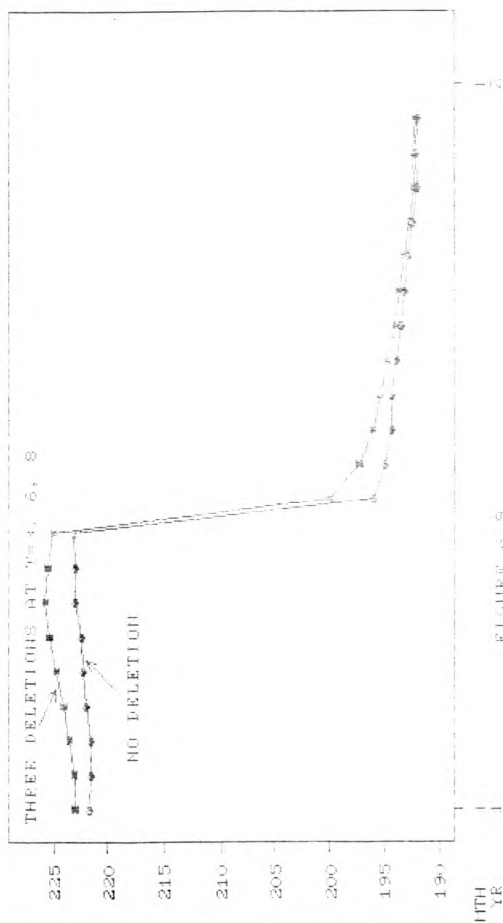


FIGURE 3.9

CHAPTER SEVEN

CUSUMS AND MODEL MONITORING - A REVIEW

7.1 Introduction

The second part of the thesis which deals with model monitoring starts with this chapter. The model monitoring mechanism studied here makes use of the cumulative sum or cusum technique, which was developed in the 1950s in the context of improving the then prevailing control mechanism for manufacturing processes.

In line with the first part of the thesis this second part deals with unusual values and changes in a given process. However, here we deal with the tracking down of outlying observations with the objective of bringing corresponding changes to the process or model under investigation.

In this chapter, the cusum is introduced and explained; key related ideas are presented. In the following chapters thereafter, new approaches to cusum as a monitor are presented.

7.2 General Background

Manufacturing processes always need to be monitored or controlled in some way for obvious economic reasons. That need gave rise to statistical quality control or statistical process control. The motivation behind a control mechanism for such a process is to obtain information on the process so as to verify whether the process

is operating in a given specified manner or to warn as quickly as possible whether some departure from specifications is occurring so that appropriate action may be taken. It is desirable that such information should be displayed with a view to making the control mechanism more effective. The display of the gathered information is expected to be such that it will readily help to identify any departure from specifications and so make it possible to isolate any causes of trouble.

The Shewhart control chart developed in the 1930s was the first to achieve these objectives but to a limited extent. The chart consists essentially of a line which indicates the level at which the process is expected to operate with action lines, usually at a distance of $\pm 3.09\sigma$ from the level. This chart was improved by including warning lines, usually at distance of $\pm 1.96\sigma$ from the level. Such a presentation would be reinforced by some rules-of-thumb for making decisions based upon the data. Various minor improvements were made to the Shewhart control charts.

However, they fail to provide a wholly satisfactory control mechanism; they tend to identify the departure from specifications rather late in time and they do not make good use of all the available data and this in spite of the fact that they effectively demonstrated how proper visual presentation of data could be very useful. Only the last value or last few values are taken into account in some way: no use whatsoever is made of the information provided by so many previous values. Even then only the zone, within which a particular value lies (e.g. outside $\pm 3.09\sigma$ limits), is considered rather than the actual numerical values obtained.

The lack of efficiency of the Shewhart control chart is largely due to the fact that its design is based on the theory of classical hypothesis testing prevailing at the time of its development. Following the introduction of sequential tests by Wald (1947) a new chart was conceived by Page (1954): the cumulative sum chart. It was further developed into a cusum decision interval scheme by Ewan and Kemp (1960). The shortcomings of the Shewhart control chart were dealt with to a large extent. All values are used in the chart and a change in the slope helps to detect more quickly any departure from specifications, especially when the change is small.

The essential feature of the cusum technique is that successive values of a variable are compared with a predetermined reference value. The cumulative sum of deviations from this value is plotted on a chart or recorded in tabulation. If the cumulation exceeds a pre-determined decision interval, this would indicate that a change has occurred in the mean level of the variable. Key characteristics of the cusum chart are: (i) the reference value, denoted usually by k , (ii) the decision interval, denoted by h , (iii) the run length, i.e. the number of points to an action signal and in particular, the average run length, denoted by ARL.

The theoretical reason why the cusum technique has proved to be so much better is that it is not based, as the Shewhart control chart, on classical hypothesis testing; rather it is perceived as a sequence of sequential tests, which help to bring in the notion of run length. There have been various developments along this line of thought brought about by Bissell (1969), Kemp (1971), Goel and Wu

(1971). Techniques based on numerical integration - difference equations were used to obtain ARL's for various control schemes.

However, as early as 1959, G. Barnard developed a new and equivalent interpretation of the Cusum technique by considering the cusum as a stochastic process: the concept of the V-mask was developed. In fact, the cusum chart and the V-mask can be used jointly and do thus provide a quite powerful technique.

The key problem, of course, was to obtain precise values of ARL for different control schemes, i.e. for different pairs of values of k and h . By 1972, Brook and Evans considered the cusum as a Markov chain, and thus they obtained not only the ARL, but the actual probability distribution of the run length. Further developments regarding the extension of the idea to two-sided cusums were brought about by Woodall (1984) and Crosier (1986).

7.3 The CUSUM

Consider a particular process with a particular characteristic of interest described by the random variable X . X may be continuous or discrete. The objective is to monitor the 'behaviour' of a parameter of interest: for example, the mean or variance. We would like to know, in particular, whether there are serious changes in the process level or major increases in variability.

Let μ be the parameter of interest. Then the problem may be formulated in terms of two simple hypotheses phrased in terms of

μ . When the process or model is under control, the parameter μ is expected to be equal to μ_0 ; when the process is out of control, $\mu = \mu_1$. μ_0 and μ_1 are predetermined values known, in quality control terms, as Acceptable Quality Level (AQL) and Rejectable Quality Level (RQL). Without loss of generality, we take $\mu_0 \leq \mu_1$. This would give us the case of one-sided cusum system; a two-sided system is obviously defined when μ_1 can be either less or greater than μ_0 and is essentially equivalent to operating two single sided schemes.

Let x_1, x_2, \dots, x_r be observations on variable X where r is called the run length. The run length is the number of observations between the time any monitoring scheme has been initialised or reinitialised and the time when a signal for action is issued by the monitoring scheme. A reference value k is defined by

$k = (\mu_1 + \mu_0)/2$ and let h , a predetermined value, be such that $h > 0$.

Then a single sided Cusum monitoring scheme involves a three point decision space:

δ_0 : process acceptable, reinitialise monitor;

δ_c : continue with a further observation;

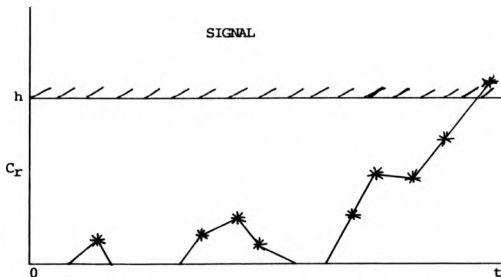
δ_1 : process questionable, issue monitor signal.

A Cusum Decision Scheme $CD(k, h)$ uses the following cumsums as monitor

$$C_r = \sum_{i=1}^r (x_i - k) = C_{r-1} + x_r - k.$$

If $C_r \geq h$, decision $\hat{\delta}_1$ is taken and if $C_r \leq 0$, decision $\hat{\delta}_0$ is taken. If either of these decisions is made then, after appropriate intervention in the case of $\hat{\delta}_1$, the scheme is reinitialised, setting $C_0 = r = 0$.

The Decision Scheme can be represented by the rather self-explanatory diagram below.



As mentioned by Page (1954), Ewan and Kemp (1960), the cusum decision scheme $CD(k, h)$ is equivalent to a succession of Wald sequential tests with horizontal boundaries distance h apart, with the lower line taken as zero. And with this in mind the average run length, ARL is computed.

Ewan and Kemp (1960) were the first to develop systematic methods to obtain ARLs for a wide variety of cusum decision schemes as well as for normal, binomial and Poisson variates. They produced special tables and special graphs known as nomograms. The nomograms give ARLs for different values of the decision interval h , of the sample size n and of the Acceptable Quality Level on the one hand and of the Rejectable Quality Level on the other hand. These nomograms are still very useful and nowadays they are widely used in quality control in manufacturing processes. Woodward and Goldsmith (1964) brought further development and clarification along the same line of thought.

A two-sided cusum decision scheme consists of operating simultaneously two single sided cusum decision schemes in which the same random variable X is used for both. Whilst the first case is as described above, in the second scheme if $C_r \leq -h$, decision δ_1 is taken and if $C_r \geq 0$, decision δ_0 is taken. Kemp (1971) showed that the ARL for such a two-sided scheme is given by

$$L^{-1} = L_1^{-1} + L_2^{-1}$$

where L_1 , L_2 are the ARL's of the separate schemes.

Finally, a last point on the cusum needs to be mentioned as it brings notation of some relevance in the discussion in the next section. As the visual impression of cusum charts is of obvious importance, the relationship between the scales used to plot C_r and r becomes an important factor. For this reason, it is recommended

that the ratio between the horizontal and vertical scales be $f:1$ where f is approximately twice the standard deviation of the plotted points.

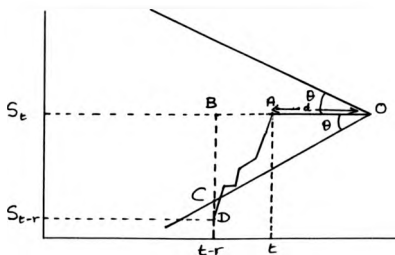
7.4 The V-Mask

G. Barnard (1959), in an attempt to provide a simpler form of a two-sided cusum scheme than that of Page (1957), developed a different way of presenting data: the V-mask Cusum graph $VG(d, \theta)$. This graph plots cumulative sums S_t against t , where

$$S_t = \sum_{j=1}^t (x_j - \mu_0) = S_{t-1} + x_t - \mu_0$$

It is to be noted that, in this case, the 'target value' or AQL μ_0 is subtracted from each value of x_i . It is expected that as long as the process parameter remains near the target value, the graph of the S_t 's should not deviate too much from the horizontal. There is a need to check this expectation.

A sheet of cardboard with a V-shaped hole cut out of it is placed on the chart with the vertex of the V pointing horizontally forwards, at a distance d ahead of the last (or leading) point on the chart and with a half-angle, θ , with the horizontal as in the diagram below.



Then if all of the data points since the last initialisation remain visible, the process is considered to be under control. If any such data point is not visible, i.e. the graph crosses either limb of the V-mask, then the process is considered to be out of control. Without loss of generality, we consider the case $\mu_0 < \mu_1$. Then decision \hat{d}_1 is made if the cusum path goes outside the lower limb of the mask,

i.e. if, from the diagram, $BD \geq BC$

i.e. if $S_t - S_{t-r} \geq (d + r) f \tan \theta$, where f is the scale factor

i.e. if $\sum_{i=t-r+1}^t (x_i - \mu_0) \geq (d + r) f \tan \theta$

i.e. if $\sum_{i=t-r+1}^t (x_i - \mu_0 - f \tan \theta) \geq f d \tan \theta$.

The V-mask Cusum graph $VG(d, \theta)$ is then equivalent to a

Cusum Decision Scheme $CD(k, h)$ if

$$k = \mu_0 + f \tan \theta \quad \text{and} \quad h = fd \tan \theta$$

$$\text{i.e. if } \tan \theta = \frac{\mu_1 - \mu_0}{2f} \quad \text{and} \quad h = \frac{d}{2} (\mu_1 - \mu_0)$$

A similar argument holds in the case of the upper limb of the V mask, so that the full V mask can be used for a double sided decision scheme; alternatively, the appropriate half of it can be used for the corresponding one sided scheme.

Though a number of modifications to the V-mask approach by using a truncated V-mask or a parabolic one, have been proposed as in Bissell (1969), the fundamental principles are basically the same.

7.5 Other Aspects of Cusums

Certain variations of the cusum technique have been proposed. Thus, for example, Lucas and Crosier (1982b) developed the idea of a robust cusum by considering contaminated normal distributions to describe the process characteristic under investigation; Lucas and Crosier (1982a) introduced the fast initial response for cusum by starting the monitor with $S_0 > 0$, so that when a process is out of control, a signal will be given faster than with the usual cusum monitor. All these modifications are useful from a practical point of view but they do not bring any new fundamental concept regarding the Cusum technique.

However, regarding the computation of the Average Run Length, Brook and Evans' (1972) work, mentioned in sub-section 7.2, is particularly useful and efficient for cusum decision scheme. The approach they used consists in considering the cusum as a Markov chain which is very different to the traditional approached used by, for example, Page (1954), Ewan and Kemp (1960), Bissell (1909), Kemp (1971). With this approach, the actual probability distribution of run length with its moments and percentage points can be obtained. And, in particular, it helps to examine the effect of departures from the assumed probability distribution so that it is useful in investigating the robustness of average run length.

Consider first the discrete case. Let k, h be as defined previously, but having positive integer values. Let the cumulative sum S_t take integer values $0, 1, \dots, h$. Then if $S_t = i$, the cusum scheme is said to be in state E_i . Each realisation of the scheme can be regarded as a random walk over the states E_0, E_1, \dots, E_h , where E_h is an absorbing state. The process is assumed to be initially in state E_0 .

Then making use of the underlying distribution for the process, the transition probabilities from state E_i can then be computed with

$$p_{i0} = P(E_i \rightarrow E_0),$$

$$p_{ij} = P(E_i \rightarrow E_j), \quad i = 1, 2, \dots, h-1,$$

$$p_{ih} = P(E_i \rightarrow E_h)$$

The transition probability matrix for the Markov chain is thus obtained, given h and k . Hence the exact probability distribution of run length and its moments can be determined.

The continuous case is dealt with by approximating the distribution by a set of class intervals which then define a finite set of states.

7.6 Cusum and Model Monitoring

The cusum technique was developed to monitor the performance of manufacturing processes by effectively monitoring the performance of the statistical model which describes the behaviour of the process. It is precisely as a general model monitor that we are here interested in the cusum technique.

The key ideas of cusum which are found so useful in model monitoring can be summarised briefly as follows.

- (i) The technique is sequential so that it helps to track down local model failure.
- (ii) The timing of model failure can be detected with some precision.
- (iii) Visual presentation of the monitor is very effective.

There are, of course, other aspects of the cusum which may not be entirely that useful in model monitoring. Usually, an

alternative model as opposed to the standard model under investigation needs to be specified. Specifically, the types of model failure need to be anticipated such as the occurrence of outliers, structural change in some variables. Also, as a result of its sequential nature, issues of masking would become relevant.

Harrison and Davies (1964) were the first to use the cusum technique to monitor forecasting systems. Cusums of the forecasting errors are calculated, appropriate limits are defined and then the usual principles of cusum are used to track down poor forecasting performance. The technique used is known as 'backward cusum' because the testing principle is based on a 'backward' sequential test.

Among later developments to model monitoring, it is worthwhile to highlight the work of West (1986a), West and Harrison (1986b, 1989) who make use of the Bayes Factor to develop an appropriate monitoring mechanism in a Bayesian perspective. Even then, the idea of a 'Bayesian cusum' is used for the monitor Z_t . In fact, $Z_t = \log(V_t)$ where

$$V_t = \min_{1 \leq k \leq t} W_t(k)$$

and $W_t(k)$ is the cumulative Bayes Factor.

CHAPTER EIGHTSEQUENCES OF SPRT's (SSPRT's)8.1 Introduction

As explained in chapter 7, a cusum monitoring scheme involves a three point decision space, with decisions usually related to two simple hypotheses phrased in terms of the parameter under scrutiny. Various authors have stressed the point that the cusum is essentially a sequence of sequential tests (e.g. Page (1954), Ewan and Kemp (1960), Bissell (1969)). And, in fact, the ideas of sequential probability ratio test (SPRT) are used to develop techniques to obtain, amongst other things, the Average Run length which is a key summary of any cusum decision scheme. But it does not seem to have occurred to the various authors that the cusum monitor itself can be perceived as a special case of a sequence of Sequential Probability Ratio Tests (SSPRT's).

In this Chapter, this is established for the exponential family of distributions.

8.2 The Sequential Probability Ratio Test (SPRT)

The SPRT is a technique based on the application of sequential methods to classical hypothesis testing. Let the hypotheses be defined as follows:

$$H_0 : \text{Model is standard with } X_1 | H_0 \sim f(\cdot | H_0)$$

H_1 : Alternative model with $X_i | H_1 \sim f(\cdot | H_1)$,

where $f(\cdot | H_0)$ and $f(\cdot | H_1)$ are of the same functional form but having different values for the parameter of interest. Further let x_1, x_2, \dots be successive observations which are outcomes of independent random variables X_i . Then for any positive integer m , the probability that a sample x_1, x_2, \dots, x_m is obtained is given by

$$P_{0m} = P(x_1, x_2, \dots, x_m | H_0) = f(x_1 | H_0) \cdot f(x_2 | H_0) \dots f(x_m | H_0) \text{ and} \\ P_{1m} = P(x_1, x_2, \dots, x_m | H_1) = f(x_1 | H_1) \cdot f(x_2 | H_1) \dots f(x_m | H_1).$$

The sequential probability ratio test for testing H_0 against H_1 is then defined as follows: two positive constants A and B (with $B < A$) are defined. At the m^{th} observation, the ratio P_{1m}/P_{0m} is computed for $m = 1, 2, \dots$. Then the following three point decision space is defined. If

$$(i) \quad B < \frac{P_{1m}}{P_{0m}} < A, \text{ continue by taking another observation;}$$

$$(ii) \quad \frac{P_{1m}}{P_{0m}} \geq A, \quad H_0 \text{ is rejected, i.e. } H_1 \text{ is accepted and} \\ \text{process is terminated;}$$

$$(iii) \quad \frac{P_{1m}}{P_{0m}} \leq B, \quad H_0 \text{ is accepted and the process is terminated.}$$

Alternatively, taking logarithms, we have

$$\begin{aligned}\log \frac{P_1 m}{P_0 m} &= \sum_{i=1}^m \log \frac{f(x_i | H_1)}{f(x_i | H_0)} \\ &= \sum_{i=1}^m z_i, \text{ say where } z_i = \log \frac{f(x_i | H_1)}{f(x_i | H_0)}\end{aligned}$$

The corresponding decision space is then defined as follows:

If

(i) $\log B < \sum_{i=1}^m z_i < \log A$, continue by taking another observation;

(ii) $\sum_{i=1}^m z_i \geq \log A$, H_0 is rejected, i.e. H_1 is accepted and process is terminated;

(iii) $\sum_{i=1}^m z_i \leq \log B$, H_0 is accepted and the process is terminated.

A and B are constants related to the values of α , the probability of rejecting H_0 when H_0 is true and β , the probability of rejecting H_1 when H_1 is true. They are the usual type I and type II errors in classical hypothesis testing.

8.3 Sequences of SPRT's (SSPRT's)

8.3.1 Definition

Consider the two hypotheses H_0 and H_1 as defined in subsection 8.2, viz. H_0 : Model is standard with $X_i | H_0 \sim f(\cdot | H_0)$
 H_1 : Alternative model with $X_i | H_1 \sim f(\cdot | H_1)$.

The monitor used here is based on the Bayes' Factor, as defined by Jeffreys (1961). The Bayes' Factor, B_t , is essentially the predictive probability ratio defined by

$$B_t = \frac{f(X_t | H_0)}{f(X_t | H_1)}.$$

It is to be noted here that once $X_t = x_t$ is observed, then the Bayes' Factor is also the usual likelihood ratio with

$$\frac{L(H_0 | x_t)}{L(H_1 | x_t)} = \frac{f(x_t | H_0)}{f(x_t | H_1)}.$$

Let r_t be the run length at time t , as defined in Chapter 7, so that $r_t = t - t_0$ where the last decision on H_0 and H_1 occurred at time t_0 . At time t , the Bayes' Factor based on the sequence of r_t observations x_{t_0+1}, \dots, x_t is defined as

$$L_t(r_t) = \prod_{i=0}^{r_t-1} B_{t-i},$$

$$\begin{aligned}
 \text{so that } L_t(r_t) &= B_t \prod_{i=1}^{r_t-1} B_{t-i} \\
 &= B_t \prod_{i=0}^{r_t-2} B_{t-1-i} \\
 &= B_t \cdot L_{t-1}(r_t - 1)
 \end{aligned} \tag{8.1}$$

For each r_t , the Bayes' Factor $L_t(r_t)$ measures the evidence provided by the most recent r_t observations for or against the standard model. In particular, it is to be noted that the evidence accumulates multiplicatively as the data are processed, as shown by equation (8.1). Further, following Jeffreys (1961), a log Bayes' factor of -1 indicates evidence in favour of the model defined by H_1 , a value of -2 indicating the evidence to be strong and a value of 0 indicating obviously no evidence either way. Positive values of 1 and 2 would give corresponding evidence in favour of the model defined by H_0 .

The characterisation of the Bayes factor mentioned above would suggest the definition of a monitor based on the negative of its logarithm. And this would give a simple cumulative procedure. Effectively, define the test quantity T_t by

$$\begin{aligned}
 T_t &= -\log L_t(r_t) \\
 &= -\sum_{i=1}^{r_t-1} \log B_{t-i}
 \end{aligned}$$

Then with $s > 1$, and $a < s$, a three point decision space would be defined as follows:

- (i) δ_0 : Accept H_0 and reinitialize monitor if $T_t \leq a$;
- (ii) δ_1 : Accept H_1 , i.e. issue monitor signal if $T_t \geq s$;
- (iii) δ_c : Continue with a further observation if $a < T_t < s$.

In cases (i) and (ii), r_{t+1} is reset to 1 and the process is restarted following any necessary intervention. Thus whilst in the case of SPRT, there is a single test which, after accepting H_0 or issuing monitor signal (i.e. accepting H_1), concludes, in this case we are concerned with the continual application of SPRT's. Once a decision has been made, another test is immediately started following any necessary process intervention. Hence we have a sequence of SPRT's, SSPRT's.

It is to be noted that, following the earlier comments on the log Bayes' factor, and based on prior information only, $T_0 = 0$, indicating no evidence either way to start with.

8.3.2 The Case of the Normal Mean (with Known Variance)

Consider the standard Normal mean SSPRT in which

$X_t | \mu \sim \text{IN}(\mu; \sigma^2)$ are independent identically distributed Normal random variables with mean μ and known variance σ^2 . In the set up of hypothesis testing, we have $H_0 : X_t | \mu_0 \sim \text{IN}(\mu_0; \sigma^2)$,
 $H_1 : X_t | \mu_1 \sim \text{IN}(\mu_1; \sigma^2)$.

The probability density functions would be defined by

$$f(x_t | H_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \left\{ \frac{(x_t - \mu_j)^2}{2\sigma^2} \right\}, \quad j = 0, 1.$$

Then, the Bayes' Factor B_t is defined by

$$\begin{aligned} B_t &= \frac{f(x_t | H_0)}{f(x_t | H_1)} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (x_t^2 - 2x_t\mu_0 + \mu_0^2 - x_t^2 + 2x_t\mu_1 - \mu_1^2) \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} [2x_t(\mu_1 - \mu_0) - (\mu_1^2 - \mu_0^2)] \right\} \\ &= \exp \left\{ -\frac{1}{\sigma^2} (\mu_1 - \mu_0) \left(x_t - \frac{\mu_1 + \mu_0}{2} \right) \right\}. \end{aligned}$$

After r_t observations, we have

$$\begin{aligned} T_t &= \sum_{i=0}^{r_t-1} -\log B_{t-i} \\ &= \frac{1}{\sigma^2} (\mu_1 - \mu_0) \sum_{i=0}^{r_t-1} (x_{t-i} - k), \end{aligned}$$

where $k = \frac{\mu_1 + \mu_0}{2}$.

The resulting three-point decision space is defined as follows:

- (i) δ_0 : Accept H_0 and reinitialize monitor if $T_t \leq a$

$$\text{i.e. if } \sum_{i=0}^{T_t-1} (x_{t-i} - k) \leq \frac{na^2}{\mu_1 - \mu_0}$$

- (ii) δ_1 : Accept H_1 , i.e. issue monitor signal if $T_t \geq s$

$$\text{i.e. if } \sum_{i=0}^{T_t-1} (x_{t-i} - k) \geq \frac{ns^2}{\mu_1 - \mu_0}$$

- (iii) δ_c : Continue otherwise.

With $T_0 = 0$ and $a = 0$, this SSPRT is clearly equivalent to a

Cusum Decision Scheme $CD(k, h)$ as defined in Chapter 7, with

$$k = \frac{\mu_1 + \mu_0}{2} \quad \text{as then defined and } h = \frac{na^2}{\mu_1 - \mu_0}.$$

8.4 The Exponential Family - A General Result

8.4.1 Definition of the Exponential Family

If X_t is assumed to have a distribution in the exponential family, then the density or probability mass function (if discrete) of X_t may be described as follows. For some defining quantities η_t and ϕ_t , and three known functions $t(X_t)$, $a(\eta_t)$

and $b(x_t, \phi_t)$, the density is

$$f(x_t | \eta_t, \phi_t) = b(x_t, \phi_t) \exp\{\phi_t[t(x_t) \eta_t - a(\eta_t)]\}, \quad (8.2)$$

over $A \subseteq R$,

where $a(\eta_t)$ is a convex twice differentiable function of η_t ,

ϕ_t is the precision parameter,

η_t is the continuous natural parameter,

$$\mu_t = E[t(x_t) | \eta_t, \phi_t] = \frac{\partial a(\eta_t)}{\partial \eta_t} = \dot{a}(\eta_t), \quad \text{with}$$

μ_t a monotonically increasing function of η_t , and

$t(x_t)$ sufficient for η given x_t .

8.4.2 The General Result

Theorem 8.1

An SSPRT with $a = 0$ for the exponential family for hypotheses

$$H_0 : x_t | \eta_0, \phi_t \sim \text{If}(\cdot | \eta_0, \phi_t)$$

$$H_1 : x_t | \eta_1, \phi_t \sim \text{If}(\cdot | \eta_1, \phi_t)$$

is equivalent to a Cusum Decision Scheme based upon

$$\sum_{i=0}^{n_t-1} (t(y_{t-i}) - k), \quad \text{with}$$

$$k = \frac{a(\eta_0) - a(\eta_1)}{\eta_0 - \eta_1} \quad \text{and} \quad h = \frac{s}{\phi_t(\eta_1 - \eta_0)}.$$

Proof:

From the definition of the exponential family, we have

$$-\log \frac{f(x_t | \eta_0, \phi_t)}{f(x_t | \eta_1, \phi_t)} = \phi_t [a(\eta_0) - a(\eta_1) + (\eta_1 - \eta_0) t(x_t)]$$

Then, the test quantity T_t is defined by

$$\begin{aligned} T_t &= \sum_{i=1}^{r_t-1} -\log B_{t-i} \\ &= r_t \phi_t [a(\eta_0) - a(\eta_1)] + \phi_t (\eta_1 - \eta_0) \sum_{i=0}^{r_t-1} t(x_{t-i}) \\ &= \phi_t (\eta_1 - \eta_0) \cdot \left\{ \sum_{i=0}^{r_t-1} \left[t(x_{t-i}) - \frac{a(\eta_0) - a(\eta_1)}{\eta_0 - \eta_1} \right] \right\} \\ &= \phi_t (\eta_1 - \eta_0) \sum_{i=0}^{r_t-1} [t(x_{t-i}) - k], \end{aligned}$$

where $k = \frac{a(\eta_0) - a(\eta_1)}{\eta_0 - \eta_1}$.

Finally from subsection 8.4.1, we have $T_t \geq s$ for issuing monitor signal, so that then

$$\sum_{i=0}^{r_t-1} [t(x_{t-i}) - k] \geq \frac{s}{\phi_t (\eta_1 - \eta_0)},$$

thus establishing the equivalence between the SSPRT and a Cusum Decision Scheme $CD(k, h)$ with k, h as defined above.

8.5 Special Cases

8.5.1 The Binomial Model

Consider the Binomial SSPRT in which $X_t | n, p \sim IB(n, p)$. Then, as in previous cases, the hypotheses are defined as follows:

$$H_0 : X_t | n, p_0 \sim IB(n, p_0)$$

$$H_1 : X_t | n, p_1 \sim IB(n, p_1)$$

In the case of the binomial model, the parameter of interest is p . In practical cases, p could be, for example, the proportion of defectives in a given batch of manufactured goods. The monitor is meant to track down any deterioration in this proportion; in particular, a signal is issued if p exceeds a certain limit, so that $p_1 > p_0$. In such cases, samples are drawn and X_t would denote the number of defectives in a sample.

By definition, $p(x | n, p_i) = \binom{n}{x} p_i^x q_i^{n-x}$, where $i = 0, 1$ and $q_i = 1 - p_i$. Then the Bayes' Factor, B_t is defined by

$$B_t = \left(\frac{p_0}{p_1} \right)^{x_t} \left(\frac{q_0}{q_1} \right)^{n-x_t} = \left(\frac{q_0}{q_1} \right)^n \left(\frac{p_0 q_1}{p_1 q_0} \right)^{x_t}.$$

$$\begin{aligned}
 \text{And } T_t &= \sum_{i=0}^{r_t-1} -\log B_{t-i} \\
 &= nr_t \log \left(\frac{q_1}{q_0} \right) - \log \left(\frac{p_0 q_1}{p_1 q_0} \right) \sum_{i=0}^{r_t-1} x_{t-i} \\
 &= n \log \left(\frac{p_1 q_0}{p_0 q_1} \right) \left[\sum_{i=0}^{r_t-1} \left(\frac{x_{t-i}}{n} - k \right) \right],
 \end{aligned}$$

$$\begin{aligned}
 \text{where } k &= \frac{\log (q_0/q_1)}{\log (p_1 q_0/p_0 q_1)} \\
 &= \frac{\log (q_1/q_0)}{\log (p_0 q_1/p_1 q_0)} \\
 &= \frac{\log q_1 - \log q_0}{\log (p_0/q_0) - \log (p_1/q_1)} \\
 &= \frac{-\log q_0 - (-\log q_1)}{\log (p_0/q_0) - \log (p_1/q_1)}
 \end{aligned}$$

$$\text{And } h = \frac{S}{n \log (p_1 q_0/p_0 q_1)} = \frac{S}{n[\log (p_1/q_1) - \log (p_0/q_0)]}$$

We note that the Binomial is a member of the exponential family with

$$\eta_t = \log \left(\frac{p_i}{1-p_i} \right) = \log \left(\frac{p_i}{q_i} \right),$$

$$\phi_t = n,$$

$$a(\eta_t) = \log (1 + \exp \eta_t)$$

$$= \log (1 + \frac{p}{q})$$

$$= - \log q ,$$

$$t(x_t) = \frac{x_t}{n} ,$$

so that, clearly, $k = \frac{a(\eta_0) - a(\eta_1)}{\eta_0 - \eta_1}$, and

$$h = \frac{g}{\theta_t(\eta_1 - \eta_0)}$$

Thus the Cusum Decision Scheme (k, h) with k and h as defined above and making use of $t(x_t) = \frac{x_t}{n}$ is equivalent to the Binomial SSPRT.

8.5.2 The Poisson Case

Consider the Poisson SSPRT in which $X_t | \theta \sim IP(\theta)$. Then the hypotheses are, as usual, defined by

$$H_0 : X_t | \theta_0 \sim P(\theta_0)$$

$$H_1 : X_t | \theta_1 \sim P(\theta_1)$$

By definition, $p(x_t/\theta_i) = \theta_i^{x_t} \exp(-\theta_i)/x_t!$, where $i = 0, 1$.

Then the Bayes Factor, $B_t = \left(\frac{\theta_0}{\theta_1}\right)^{x_t} \exp. (\theta_1 - \theta_0)$

$$\begin{aligned}
 \text{Therefore } T_t &= -\log\left(\frac{\theta_0}{\theta_1}\right) \sum_{i=0}^{r_t-1} x_{t-i} - r_t(\theta_1 - \theta_0) \\
 &= \log\left(\frac{\theta_1}{\theta_0}\right) \sum_{i=0}^{r_t-1} (x_{t-i} - k)
 \end{aligned}$$

$$\text{with } k = \frac{\theta_1 - \theta_0}{\log \theta_1 - \log \theta_0} = \frac{\theta_0 - \theta_1}{\log \theta_0 - \log \theta_1}$$

$$\text{and } h = \frac{s}{\log \theta_1 - \log \theta_0} .$$

Bearing in mind that the Poisson model is a member of the exponential family with

$$\eta_t = \log \theta , \quad a(\eta_t) = \theta = \exp \eta , \quad \theta_t = 1 ,$$

it is then obvious again that $k = \frac{a(\eta_0) - a(\eta_1)}{\eta_0 - \eta_1}$ and $h = \frac{s}{\eta_1 - \eta_0}$. Thus, the Cusum Decision Scheme $CD(k, h)$ with k, h as defined above is equivalent to the SSPRT for Poisson models.

8.5.3 The Normal Variance Case

So far the cases concerning the mean have been dealt with; in this subsection we consider the monitoring mechanism for variability in a normal model. More specifically, increase in variance would be the typical case. The objective is to track down

when the variance changes from an accepted value σ^2 to a rejectable value $k\sigma^2$ with $k > 1$ or σ^2/γ with $0 < \gamma < 1$. The hypotheses are formulated as follows, with the usual notation.

$$H_0 : x_t | H_0 \sim \text{IN}(\mu ; \sigma^2)$$

$$H_1 : x_t | H_1 \sim \text{IN}(\mu ; \sigma^2/\gamma) ,$$

where μ , σ^2 and γ are known.

In this case, we consider for, obvious reasons, samples of values at a time as opposed to single values as in the cases considered so far. Thus, the hypotheses could be then reformulated, with n , known sample size.

$$H_0 : \bar{x}_t | H_0 \sim \text{IN}(\mu ; \sigma^2/n)$$

$$H_1 : \bar{x}_t | H_1 \sim \text{IN}(\mu ; \sigma^2/\gamma n)$$

writing $\gamma_0 = 1$, and $\gamma_1 = \gamma$, the density functions are then given by

$$\begin{aligned} f(\bar{x}_t | H_i) &= \sqrt{\frac{n\gamma_i}{2\pi\sigma^2}} \exp \left[\frac{-n\gamma_i}{2\sigma^2} (\bar{x} - \mu)^2 \right] , \quad i = 0, 1 \\ &= \sqrt{\frac{n}{2\pi\sigma^2}} \exp \log \gamma_i \cdot \exp \left[-\frac{n\gamma_i}{2\sigma^2} (\bar{x} - \mu)^2 \right] \\ &= \sqrt{\frac{n}{2\pi\sigma^2}} \exp \left[\frac{n}{2\sigma^2} [t(x) \eta - a(\eta)] \right] \end{aligned}$$

$$\text{where } t(x) = (\bar{x} - \mu)^2,$$

$$\eta = -Y_i,$$

$$\phi_t = \left(\frac{2\theta^2}{n} \right)^{-1},$$

$$a(\eta) = -\frac{n}{2} \log(-\eta) = -\frac{n}{2} \log(Y_i).$$

This shows clearly that we are dealing with a member of the Exponential Family. Theorem 8.1 is now proved to be valid in this case as well.

We have

$$\begin{aligned} \log B_t &= \log \frac{f(\bar{x}_t | H_0)}{f(\bar{x}_t | H_1)} \\ &= -\frac{1}{2} \log Y_0 - \frac{nY_0}{2\theta^2} (\bar{x}_t - \mu)^2 - \frac{1}{2} \log Y + \frac{nY}{2\theta^2} (\bar{x}_t - \mu)^2 \\ &= -\frac{1}{2} \log Y - \frac{n}{2\theta^2} (1 - Y) (\bar{x}_t - \mu)^2, \text{ since } Y_0 = 1, Y_1 = Y \end{aligned}$$

$$\begin{aligned} \text{Then } T_t &= - \sum_{i=0}^{r_t-1} \log B_{t-i} \\ &= \frac{r_t}{2} \log Y + \frac{n}{2\theta^2} (1 - Y) \sum_{i=0}^{r_t-1} (\bar{x}_{t-i} - \mu)^2 \\ &= \frac{n}{2\theta^2} (1 - Y) \sum_{i=0}^{r_t-1} (\bar{x}_{t-i} - \mu)^2 = k, \end{aligned}$$

$$\begin{aligned}
 \text{where } k &= -\frac{g^2 \log Y}{n(1-Y)} \\
 &= \frac{-(-\frac{g^2}{n} \log Y_0) + (-\frac{g^2}{n} \log Y)}{-Y - (-Y_0)}, \text{ with } Y_0 = 1 \\
 &= \frac{-a(\eta_0) + a(\eta_1)}{\eta_1 - \eta_0} \\
 &= \frac{a(\eta_0) - a(\eta_1)}{\eta_0 - \eta_1}.
 \end{aligned}$$

And finally, h is defined by $h = \frac{g^2 n^2}{n(1-Y)}$

$$\begin{aligned}
 \text{i.e. } h &= \frac{g^2}{\frac{g^2}{h} - 1(1-Y)} \\
 &= \frac{g^2}{g^2(\eta_1 - \eta_0)}.
 \end{aligned}$$

Thus, the Cusum Decision Scheme $CD(k, h)$ with k, h as defined above is equivalent to the SSPRT for the normal variance case.

8.5.4 The Case of t-distribution

So far, we have been considering models which belong to the exponential family whereby the equivalence between the Cusum Decision scheme and the SSPRTs is established. We now consider the t -distribution which is not a member of the exponential family; the

SSPRTs exists but there does not exist an equivalent Cusum.

Consider the variable X_t having a t-distribution, so that the hypotheses are defined by

$$H_0 : X_t \mid H_0 \sim T_{n_t} [f_t ; Q_t]$$

$$H_1 : X_t \mid H_1 \sim T_{n_t} [f_t + \mu ; Q_t/\gamma] ,$$

where n_t = number of degrees of freedom, f_t = location parameter and Q_t = scale parameter.

Then the probability density function of X_t is defined by

$$p(x_t \mid H_0) = \frac{\Gamma[(n_t + 1)/2]}{(Q_t n_t)^{1/2} \Gamma(n_t/2)} \left[1 + \frac{(x_t - f_t)^2}{Q_t n_t} \right]^{-(n_t + 1)/2}$$

Thus, we have

$$2 \log p(x_t \mid H_0) = -\log Q_t - (n_t + 1) \log \left[\frac{n_t Q_t + e_t^2}{n_t Q_t} \right] + \text{constant} ,$$

where $e_t = x_t - f_t$.

Similarly, we have

$$2 \log p(x_t \mid H_1) = \log \gamma - \log Q_t - (n_t + 1) \log \left[\frac{n_t Q_t + \gamma(e_t - \mu)^2}{n_t Q_t} \right] + \text{constant} .$$

Then from $T_t = - \sum_{i=0}^{r_t-1} \log B_{t-i}$, we have

$$T_t = \frac{1}{2} \left[r_t \log \gamma + \sum_{i=0}^{r_t-1} (n_{t-i} + 1) \log \left(\frac{n_{t-i} Q_{t-i} + e_{t-i}^2}{n_{t-i} Q_{t-i} + \gamma (e_{t-i} - \mu)^2} \right) \right]$$

Then the SSPRT is defined using the decision space defined in Section 8.3.1. We have

- (i) δ_0 : Accept H_0 and reinitialize monitor if $T_t \leq a$.
- (ii) δ_1 : Accept H_1 , i.e. issue monitor signal if $T_t \geq s$.
- (iii) δ_c : Continue with a further observation if $a < T_t < s$.

It is to be noted that we do not get an expression for T_t which is a cumulative sum as should normally be expected, but where $t(x)$, a sufficient statistic for x is not present as in the case of exponential family models. Thus there is no equivalence here between the SSPRT and the Cusum Decision scheme.

Finally, in this case, there are in fact four alternatives to the standard model depending upon where the emphasis lies. The four alternatives that might be tested are given by

- (i) $Y = 1$, and $\mu = \mu_1 > 0$, so that here we are monitoring an increase in level.
- (ii) $Y = 1$, and $\mu = \mu_2 < 0$, so that here we are monitoring a decrease in level.
- (iii) $\mu = 0$ and $Y = Y_1 > 1$, monitoring an increase in variability.
- (iv) $\mu = 0$ and $Y = Y_2 < 1$, monitoring the unlikely case of decrease in variability.

In each of the above four cases, the monitor T_t would be correspondingly simplified, though not in a fundamental way.

8.6 Comments

Following the comment made in Section 8.3.1 about the fact that, once $X_t = x_t$ is observed, then the Bayes' Factor is also the likelihood ratio, it is then evident that SSPRTs could be based in an equally efficient way on the posterior likelihood ratio. And this gives an advantage because then the procedures discussed here would be invariant under reparameterisation. When we adopt the procedures involving Bayes' Factor, hence probability distribution, a Jacobian ratio adjustment would be required.

CHAPTER NINE

A BAYESIAN DECISION APPROACH TO CUSUMS

9.1 Introduction

In this chapter, a Bayesian decision approach to cusums is developed. Given the coherence and flexibility characteristic of Bayesian analysis, it is not surprising to find out that this Bayesian approach offers a greater scope for generalisation. In fact, general results are obtained for the exponential family and the t -distribution.

There are two key ideas which need to be dealt with properly in order that the Bayesian approach for such sequential procedures may be successful. The first is the definition of an appropriate loss function and the second is the definition of useful and meaningful moments for the prior distribution. The next section deals mainly with the loss function.

9.2 The Loss Function

In sequential tests the expected additional information from a further observation declines as total information increases. Given initial information I_0 , let n_0 be the prior precision, in terms of equivalent observations based on I_0 . Let I_r be the information based upon the current run length r and let

$$n_r = n_0 + r.$$

It is appropriate at this stage to recall the three point decision space which characterises the cusum decision scheme $CD(k, h)$.

- δ_0 : process/model acceptable, reinitialize monitor;
- δ_C : continue with a further observation; and
- δ_1 : process/model questionable, issue monitor signal.

We note firstly that a loss function is often well represented locally by a quadratic function, i.e. the 'squared-error loss' is quite appropriate. Considering single-sided scheme, with $\mu_0 < \mu_C < \mu_1$, we can have the following loss functions for δ_1 and δ_C

- (i) $l(\delta_1, \mu | I_T) = (\mu_1 - \mu)^2$
- (ii) $l(\delta_C, \mu | I_T) = (\mu_C - \mu)^2$

The decision scheme, in fact, implies that, at each given point in time, we either stop sampling/intervene or continue taking a further observation. In the first case, i.e. stop sampling, then either decision δ_1 or δ_0 is made; in the second case, decision δ_C is made. We are thus concerned much more with consequences of decision δ_1 or δ_0 relative to δ_C and hence with their expected losses. Thus the loss function for δ_1 relative to δ_C is obtained by considering the difference between the two loss functions. Then the loss function for δ_1 relative to δ_C is given by

$$\begin{aligned} l(\delta_1, \mu | I_T) &= (\mu_1 - \mu)^2 - (\mu_C - \mu)^2 \\ &= \mu_1^2 - \mu_C^2 - 2\mu(\mu_1 - \mu_C) \end{aligned}$$

$$= \frac{\mu_1 - \mu_C}{2} \left[\frac{\mu_1 + \mu_C}{2} - \mu \right],$$

Then the relative loss function for δ_1 can be represented by

$$l(\delta_1, \mu | I_r) = a - \mu$$

so that $l(\delta_1, \mu | I_r) = c_1(a - \mu)$, with the positive proportional constant c_1 and $a > 0$, i.e. the real loss function is linear in μ .

For the loss function for decision δ_0 , the squared-error loss can be used again but with the addition of a further term which tends to zero as the posterior precision (i.e. n_r) increases. The reason for this term is that the expected information of a further observation relative to the information currently contained in the run declines with n_r . Also, as n_r increases, the probability that the process goes out of control within the run increases. Such a change may then be belatedly detected, whereas a new run will be expected to quickly detect the lack of control. Consequently the utility of run length decreases with n_r . Thus the loss function for δ_0 can be represented by

$l(\delta_0, \mu | I_r) = (\mu_0 - \mu)^2 + \frac{2b}{n_r} (\mu_C - \mu_0)$ and the loss function for δ_0 relative to δ_C can be represented by

$$l(\delta_0, \mu | I_r) = (\mu_0 - \mu)^2 + \frac{2b}{n_r} (\mu_C - \mu_0) - (\mu_C - \mu)^2$$

$$\begin{aligned}
&= 2\mu(\mu_C - \mu_O) - (\mu_C^2 - \mu_O^2) + \frac{2b}{n_r}(\mu_C - \mu_O) \\
&= 2(\mu_C - \mu_O) \left[\mu - \frac{\mu_C + \mu_O}{2} + \frac{b}{n_r} \right]
\end{aligned}$$

Hence the relative loss function for δ_O can be represented by

$$l(\delta_O, \mu | I_r) = \mu - a + \frac{b}{n_r}, \quad \text{with } a > 0 \quad (9.2)$$

so that we can write

$$l(\delta_O, \mu | I_r) = c_0(\mu - a + \frac{b}{n_r}),$$

with c_0 positive, since $\mu_C > \mu_O$.

It is to be noted that, generally, the relative loss function can be expressed in linear form using Taylor series expansion without having to define the loss functions for the δ_i 's, $i = 0, 1$, as done above.

The above relative loss functions are clearly monotonic functions of μ , decreasing for δ_1 and increasing for δ_O . Further, given that $E(\mu | I_r) = m_r$ exists, then clearly, for $i = 0$ and 1 , δ_i is the unique Bayes' decision if the corresponding posterior expected relative loss

$$E[l(\delta_i, \mu | I_r)] \leq 0 \quad (9.3)$$

If both these posterior expected relative losses are positive then $\hat{\theta}_C$ is the unique Bayes' decision.

9.3 The Normal Mean Case

9.3.1 The Model

Consider the sampling normal model in which $X_i | \mu \sim \text{IN}(\mu; \sigma^2)$ are independent identically distributed Normal random variables with mean μ and known variance σ^2 . Without loss of generality, we can take $\sigma^2 = 1$.

9.3.2 Learning on y

Since at the start of each run the process is assumed to be in control, i.e. the standard model holds, the prior distribution for μ based upon I_0 usually has mean equal to μ_0 and a corresponding precision equivalent to n_0 observations as mentioned in Section 9.2, so that $\mu | I_0 \sim N(\mu_0; 1/n_0)$.

Using Bayes' theorem and standard conjugate analysis, the posterior distribution for μ after r points in a run is obtained as follows.

$$P(\mu | I_r) = P(\mu | X_1, \dots, X_r, I_0)$$

= Likelihood \times Prior

$$= \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2} \sum_{i=1}^r (x_i - \mu)^2 \times \frac{\sqrt{n_0}}{\sqrt{2\pi}} \exp - \frac{1}{2} [(\mu - \mu_0) \sqrt{n_0}]^2$$

$$\begin{aligned}
&= \exp - \frac{1}{2} \left[\sum_{i=1}^r (y_i - \mu)^2 + n_0 (\mu - \mu_0)^2 \right] \\
&= \exp - \frac{1}{2} \left[\mu^2 (n_0 + r) - 2 \left(\sum_{i=1}^r y_i + n_0 \mu_0 \right) \mu \right] \\
&= \exp - \frac{1}{2} (n_0 + r) \left[\mu - \frac{\sum_{i=1}^r y_i + n_0 \mu_0}{n_0 + r} \right]^2 \\
&= \exp - \frac{1}{2} n_r [\mu - m_r]^2,
\end{aligned}$$

where $n_r = n_0 + r$

and
$$m_r = \frac{n_0 \mu_0 + \sum_{i=1}^r y_i}{n_r}$$

Thus, $(\mu | I_r) \sim N[m_r, \frac{1}{n_r}]$.

9.3.3 The Bayesian Decision Scheme BD(a, b)

Let E_1, E_0 denote the posterior expected losses for δ_1 and δ_0 respectively relative to δ_C . Then we have

$$\begin{aligned}
E_1 &= E[l(\delta_1, \mu | I_r)] \\
&= \int l(\delta_1, \mu | I_r) p(\mu | I_r) d\mu
\end{aligned}$$

$$= \int c_1(a - \mu) p(\mu | I_r) d\mu, \text{ using equation (9.1)}$$

$$= c_1 a - c_1 \int \mu p(\mu | I_r) d\mu$$

$$= c_1(a - m_r)$$

$$= c_1 \left(a - \frac{n_0 \mu_0 + \sum_{i=1}^r x_i}{n_r} \right)$$

$$\text{i.e.} \quad E_1 = E(\delta_1, m_r | I_r) = a - m_r.$$

Then using expression (9.3), decision δ_1 is taken if $E_1 \leq 0$

$$\text{i.e.} \quad \text{if } a - \frac{n_0 \mu_0 + \sum_{i=1}^r x_i}{n_0 + r} \leq 0$$

$$\text{i.e.} \quad \text{if } -n_0 \mu_0 + n_0 a + r a - \sum_{i=1}^r x_i \leq 0$$

$$\text{i.e.} \quad \text{if } n_0 a - n_0 \mu_0 - \sum_{i=1}^r (x_i - a) \leq 0$$

$$\text{i.e.} \quad \text{if } \sum_{i=1}^r (x_i - a) \geq n_0 a - n_0 \mu_0.$$

Similarly, we have

$$\begin{aligned}
 E_0 &= E\{l(\delta_0, \mu \mid I_r)\} \\
 &= \int l(\delta_0, \mu \mid I_r) p(\mu \mid I_r) d\mu \\
 &= \int c_0(\mu - a + \frac{b}{n_r}) p(\mu \mid I_r) d\mu \\
 &= c_0(\frac{b}{n_r} - a) + c_0 m_r
 \end{aligned}$$

$$\text{i.e.} \quad E_0 = l(\delta_0, m_r \mid I_r) = m_r - a + \frac{b}{n_r}$$

$$= c_0 \left(\frac{b}{n_r} - a + \frac{n_0 \mu_0 + \sum_{i=1}^r x_i}{n_0 + r} \right)$$

Then decision δ_0 is taken if $E_0 \leq 0$,

$$\text{i.e.} \quad \text{if } \frac{b}{n_0 + r} - a + \frac{n_0 \mu_0 + \sum_{i=1}^r x_i}{n_0 + r} \leq 0,$$

$$\text{i.e.} \quad \text{if } b - a n_0 - a r + n_0 \mu_0 + \sum_{i=1}^r x_i \leq 0,$$

$$\text{i.e.} \quad \text{if } n_0 \mu_0 - a n_0 + b + \sum_{i=1}^r (x_i - a) \leq 0$$

$$\text{i.e.} \quad \text{if } \sum_{i=1}^r (x_i - a) \leq a n_0 - n_0 \mu_0 - b.$$

And, finally, if otherwise, decision δ_C is taken. We have thus the Bayesian Decision Scheme, denoted by $BD(a, b)$ because the scheme is defined by the pair (a, b) .

The equivalence with the Cusum Decision Scheme (k, h) is obvious with

$$a = k = (\mu_1 + \mu_0)/2 ; b = h = n_0 a - n_0 \mu_0 = n_0 (\mu_1 - \mu_0)/2.$$

In particular we have $\mu_1 = \mu_0 + \frac{2b}{n_0}$ where $\frac{2b}{n_0}$ may be thought of as the prior value of sampling the first observation in the run.

The equivalence with the V-mask Graph $VG(d, \theta)$ is equally obvious with

$$d = n_0 \quad \text{and} \quad \tan \theta = \frac{b}{n_0 I},$$

using expressions 7.1 of subsection 7.4. Again here we have an interesting and intuitive result that $d = n_0$, the prior precision in terms of observation equivalence.

9.4 Bayesian Decision Scheme : A General Result

9.4.1 Introduction

The derivation of the Bayesian Decision Scheme for the normal mean case gives a good insight into a generalisation of the result. The key point is that, given the relative loss function for θ_0 and θ_1 , then the expressions obtained for the posterior expected relative losses depend on the posterior probability density function (or mass function) only through its mean, m_r . It seems that provided the posterior mean, m_r , is of a certain given form, then the particular result may be applicable for the relevant model.

We prove this formally in the next section. Though the proof is fairly straightforward and is along the same lines as the derivation of the result for the normal mean case, it is nevertheless very useful given that it has a quite wide range of applicability.

9.4.2 General Result

Let X_1, \dots, X_r be random variables which satisfy the following condition, with location parameter μ . μ has a prior distribution with mean μ_0 and precision equivalent to n_0 x observations, so that the posterior mean, m_r , exists and is defined by $m_r = E(\mu | I_r) = (n_0\mu_0 + \sum_{i=1}^r x_i)/n_r$ where $n_r = n_0 + r$, i.e. $n_r m_r = n_0\mu_0 + \sum_{i=1}^r x_i$. Define the decision space as usual, i.e. $\{\theta_1, \theta_0, \theta_C\}$ and let the relative loss functions of

decisions δ_i , $i = 0, 1$ with respect to δ_C be as in Section 9.2.

Thus we have

$$l(\delta_1, \mu | I_r) = a - \mu,$$

and

$$l(\delta_0, \mu | I_r) = \mu - a + \frac{b}{n_r},$$

where $a, b > 0$ and the constants of proportionality are greater than zero.

Then the Bayesian Decision Scheme, $BD(a, b)$, is equivalent to

- (i) a Cusum Decision Scheme, $CD(a, b)$ with Acceptable Quality level, AQL , $\mu_0 = a - \frac{b}{n_0}$ and Rejectable Quality Level, $\mu_1 = a + \frac{b}{n_0}$.
- (ii) a V-mask Graph, $VG(d, \theta)$ with $d = n_0$ and $\theta = \tan^{-1} \left(\frac{b}{n_0 f} \right)$

Proof:

Given I_r , the posterior expected losses for δ_1 and δ_0 relative to δ_C are as follows.

$$\begin{aligned} E_1 &= E[l(\delta_1, \mu | I_r)] \\ &= \int l(\delta_1, \mu | I_r) p(\mu | I_r) d\mu \end{aligned}$$

$$\begin{aligned}
&= \int c_1(a - \mu_1) p(\mu | I_r) d\mu, \text{ using the defined loss function} \\
&= c_1 a - c_1 \int \mu p(\mu | I_r) d\mu \\
&= c_1(a - m_r) \\
&= c_1 \left(a - \frac{n_o \mu_o + \sum_{i=1}^r x_i}{n_o + r} \right).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
E_o &= c_o(m_r - a + b/n_r) \\
&= c_o \left(\frac{b}{n_o + r} - a + \frac{n_o \mu_o + \sum_{i=1}^r x_i}{n_o + r} \right)
\end{aligned}$$

Now, the unique Bayes decision is obtained when $E_1 \leq 0$. Thus, using the results of Section 9.3.2, we have

$$(i) \quad \text{decision } d_1 \text{ is taken if } an_o - n_o \mu_o - \sum_{i=1}^r (x_i - a) \leq 0,$$

$$\text{i.e.} \quad \text{if } \sum_{i=1}^r (x_i - a) \geq an_o - n_o \mu_o.$$

$$\begin{aligned}
(ii) \quad &\text{decision } d_o \text{ is taken if } n_o \mu_o - n_o a + b + \sum_{i=1}^r (x_i - a) \leq 0 \\
\text{i.e.} \quad &\text{if } \sum_{i=1}^r (x_i - a) \leq n_o a - n_o \mu_o - b
\end{aligned}$$

(iii) decision δ_c is taken otherwise.

It is then obvious that this decision scheme is equivalent to a Cusum Decision Scheme (k, h) where

(i) $a = k = (\mu_1 + \mu_0)/2$, by definition of k ,

(ii) $h = a n_0 - n_0 \mu_0$

(iii) $n_0 a - n_0 \mu_0 - b = 0$, so that
 $b = n_0 a - n_0 \mu_0 = h$.

The last equation further gives

$$\mu_0 = a - \frac{b}{n_0}$$

This, in turn, simplifies to $\mu_0 = \frac{\mu_1 + \mu_0}{2} - \frac{b}{n_0}$, so that

$$\begin{aligned} \frac{\mu_1}{2} &= \frac{\mu_0}{2} + \frac{b}{n_0} \\ &= \frac{a}{2} - \frac{b}{2n_0} + \frac{b}{n_0} \\ &= \frac{a}{2} + \frac{b}{2n_0}. \end{aligned}$$

Hence $\mu_1 = a + \frac{b}{n_0}$, establishing thus the equivalence with the

Cusum Decision Scheme (a,b) with μ_0, μ_1 defined in terms of a,b and n_0 as required.

Given that the Cusum Decision Scheme (a,b) and the V-mask Graph $VG(d,\theta)$ are equivalent, with $\tan \theta = \frac{\mu_1 - \mu_0}{2f}$ and $b = \frac{d}{2} (\mu_1 - \mu_0)$ as per Chapter 7, then the equivalence between the Bayesian Decision Scheme $BD(a,b)$ and the V-mask Graph $VG(d,\theta)$ is obvious. As $\mu - \mu_0 = \frac{2b}{n_0}$ we have the particularly interesting result that $d = n_0$ and $\theta = \tan^{-1} \left(\frac{b}{n_0 f} \right)$. An important feature of the V-mask would be that the distance of the vertex of the V-mask from the latest plotted point is n_0 which is the equivalent number of x observations of the prior precision.

Another sensible observation is that from the definition of the relative loss function for δ_0 , i.e. $l(\delta_0, \mu | I_T) = \mu - a + \frac{b}{n_T}$, we can now see that

$$\begin{aligned} l(\delta_0, \mu | I_0) &= \mu - a + \frac{b}{n_0} \\ &= \mu - \mu_0. \end{aligned}$$

This, of course, tends to confirm both the intuitive and theoretical basis of the results obtained.

9.4.3 Comments

The concept of prior precision being equivalent to n_0 observations is an interesting and intuitive idea; it can be referred

to as 'observation precision'. The expression for the posterior mean m_r with

$$m_r = \frac{n_0 \mu_0 + \sum_{i=1}^r x_i}{n_0 + r}$$

conveys clearly this idea. m_r is, in fact, the weighted average of the data and n_0 'equivalent observations'. This idea holds for any appropriate distribution under consideration for which the general result is valid.

In the case of the normal mean discussed in Section 9.3, the prior precision n_0 is equal to the reciprocal of the variance; but generally it does not have to. In fact, in the cases discussed in Chapter 10, n_0 is not the reciprocal of the variance.

It is a different way of perceiving precision (and, more generally, variability); it brings a unified idea of precision in a general sense, somewhat along De Finetti's idea that there is no need for parametric models and that observables matter most.

CHAPTER TENSPECIAL CASES AND APPLICATIONS10.1 Introduction

In the detailed discussion of the monitoring of models using the Bayesian Decision Scheme in Chapter 9, the salient features consist in defining the parameters of the prior distribution such that their mean and precision are in the form required in the general result. The prior mean should be μ_0 and the precision equivalent to n_0 observations. As all the cases considered in this chapter are examples of well known standard conjugate families, the posterior mean in each case exists; what is therefore required is to establish that they are of the appropriate form as defined in the general result obtained in Chapter 9.

Each of the important cases considered in this Chapter will be dealt with according to this approach.

10.2 The Binomial Case

Consider the Binomial model in which $Y_i | \mu \sim IB(n; \mu)$ are independent identically distributed binomial random variables with parameter μ being monitored and with known n . Given that Y_i/n is a sufficient statistic for μ , then the monitor will be defined in terms of $X_i = Y_i/n$.

The prior distribution would be the Beta distribution with

parameters $nn_0\mu_0$ and $nn_0(1 - \mu_0)$,

i.e. $\mu | I_0 \sim \delta\{nn_0\mu_0; nn_0(1 - \mu_0)\}$. Then the prior mean is given by

$$\frac{nn_0\mu_0}{nn_0\mu_0 + nn_0(1 - \mu_0)} = \mu_0$$

and the prior variance is given by

$$\frac{nn_0\mu_0 \cdot nn_0(1 - \mu_0)}{(nn_0)^2 \cdot (nn_0 + 1)}$$

i.e.
$$\frac{\mu_0(1 - \mu_0)}{nn_0 + 1}$$

And the prior precision =
$$\frac{nn_0 + 1}{\mu_0(1 - \mu_0)}$$

$$= \frac{n}{\mu_0(1 - \mu_0)} \left(n_0 + \frac{1}{h}\right)$$

Following standard conjugate analysis, we have the posterior distribution defined as follows

$$p(\mu | I_r) = \text{likelihood} \times \text{prior}$$

$$= \mu^{\sum_{i=1}^r y_i} (1 - \mu)^{nr - \sum_{i=1}^r y_i} \times \mu^{k-1} (1 - \mu)^{l-1}$$

where $k = nn_0\mu_0$, $l = nn_0(1 - \mu_0)$

$$\mu = \frac{k + \sum_{i=1}^r y_i - 1}{(1 - \mu)} \quad \frac{l + nr - \sum_{i=1}^r y_i - 1}{(1 - \mu)}$$

Thus the posterior distribution for μ is Beta with parameters

$$\begin{aligned} k_{\text{post}} &= k + \sum_{i=1}^r y_i \\ &= nn_0\mu_0 + n \sum_{i=1}^r x_i \\ &= u_r, \text{ say} \end{aligned}$$

$$\begin{aligned} \text{and } l_{\text{post}} &= l + nr - \sum_{i=1}^r y_i \\ &= nn_0 - nn_0\mu_0 + nr - n \sum_{i=1}^r x_i \\ &= nn_0 + nr - (nn_0\mu_0 + n \sum_{i=1}^r x_i) \\ &= n(n_0 + r) - u_r \\ &= nn_r - u_r \end{aligned}$$

Therefore the posterior mean, $m_r = \frac{k_{\text{post}}}{k_{\text{post}} + l_{\text{post}}}$

$$\begin{aligned}
 &= \frac{u_r}{n n_r} \\
 &= (n_o u_o + \sum_{i=1}^r x_i) / n_r,
 \end{aligned}$$

so that $n_r m_r = n_o u_o + \sum_{i=1}^r x_i$ which is thus exactly of the same form as that encountered for the general result in Chapter 9.

It follows that, with the loss functions as defined previously in Section 9.4.2, the Bayesian Decision Scheme, $BD(a,b)$ is equivalent to the Cusum Decision Scheme (a,b) and V-mask Graph $\left(n_o, \tan^{-1} \left(\frac{b}{n_o f} \right) \right)$.

From a practical point of view, this procedure can be applied in a manufacturing process where r samples of fixed size n are taken over time. Y_i would denote the number of defectives in a given sample with μ being the probability that an item is defective. Then X_i would represent the fraction of defectives in a given sample and the monitor based on $\sum_{i=1}^r (x_i - a)$ would help to track down values of X_i which go outside specified limits.

10.3 The Poisson Case

Consider the Poisson model in which $X_i | \mu \sim I \text{ Poisson}(\mu)$ are independent identically distributed poisson random variables with parameter μ being monitored. The prior distribution would be the gamma distribution with parameters $n_0 \mu_0$ and n_0 , i.e. $\mu | I_0 \sim \Gamma(n_0 \mu_0; n_0)$. The prior mean is defined by $n_0 \mu_0 / n_0 = \mu_0$ and the prior variance is defined by $n_0 \mu_0 / n_0^2 = \mu_0 / n_0$. Then the prior precision would be equal to n_0 / μ_0 .

Following standard conjugate analysis, we have the posterior distribution defined as follows

$$p(\mu | I_r) = \text{likelihood} \times \text{prior}$$

$$= \frac{e^{-r\mu} \mu^{\sum_{i=1}^r x_i}}{\prod_{i=1}^r x_i!} \times \mu^{n_0 \mu_0 - 1} e^{-n_0 \mu}$$

$$= \mu^{n_0 \mu_0 + \sum_{i=1}^r x_i - 1} e^{-\mu(n_0 + r)}$$

Thus we have a gamma distribution with parameters $n_0 \mu_0 + \sum_{i=1}^r x_i$

and $n_0 + r$, as usual.

The posterior mean, $m_r = (n_0 \mu_0 + \sum_{i=1}^r x_i) / (n_0 + r)$,

i.e. $n_r m_r = n_0 \mu_0 + \sum_{i=1}^r x_i$, with $n_r = n_0 + r$,

which is thus exactly of the same form as that defined in the general result of Chapter 9.

The equivalence with the Cusum Decision Scheme (a,b) follows immediately, provided the same loss functions as defined in Section 9.4.2 are being used.

10.4 The Gamma Case and the Normal Variance

10.4.1 The Gamma Model

Consider the Gamma model in which $X_i | \mu \sim \Gamma(v/2; v/2\mu)$ are independent identically distributed random variables with mean, μ equal to the ratio of the two parameters. The prior distribution for μ is an Inverse Gamma distribution, so that

$$\mu^{-1} | I_0 \sim \Gamma(k_0, l_0),$$

where $k_0 = 1 + \frac{1}{2}vn_0$, and $l_0 = \frac{1}{2}vn_0\mu_0$. We shall have again in

this case standard conjugate analysis. It is appropriate to obtain the first two moments for the distribution of μ given the above prior for μ^{-1} . For this we need the following result.

Let $Y \sim \Gamma(\alpha, \beta)$. Then, the j^{th} moments are defined by

$$\begin{aligned} E(Y^j) &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{j+\alpha-1} \exp(-\beta y) dy \\ &= \frac{\Gamma(\alpha+j)}{\beta^j \Gamma(\alpha)} \int_0^\infty \frac{\beta^{j+\alpha}}{\Gamma(\alpha+j)} y^{j+\alpha-1} \exp(-\beta y) dy \\ &= \frac{\Gamma(\alpha+j)}{\beta^j \Gamma(\alpha)} \end{aligned}$$

Let $j = -1$, so that $Y^{-1} = Z$. Then, we have

$$E(Z) = E(Y^{-1}) = \frac{\Gamma(\alpha-1)}{\Gamma(\alpha)} \cdot \beta = \frac{\beta}{\alpha-1}, \quad (10.1)$$

$$\text{and} \quad E(Z^2) = E(Y^{-2}) = \frac{\Gamma(\alpha-2)}{\Gamma(\alpha)} \cdot \beta^2 = \frac{\beta^2}{(\alpha-1)(\alpha-2)},$$

$$\text{so that} \quad \text{Var}(Z) = \frac{\beta^2}{(\alpha-1)(\alpha-2)} - \frac{\beta^2}{(\alpha-1)^2} = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}.$$

Therefore, prior mean for $\mu = \frac{1}{2} v_{n_0} \mu_0 / \frac{1}{2} v_{n_0} = \mu_0$,

and prior variance for $\mu = (\frac{1}{2}vn_0\mu_0)^2 / (\frac{1}{2}vn_0)^2 (\frac{1}{2}vn_0 - 1) = \mu_0^2 / (\frac{1}{2}vn_0 - 1)$,
 so that prior precision for $\mu = (\frac{1}{2}vn_0 - 1) / \mu_0^2$.

To obtain the posterior mean m_T for μ , we follow standard conjugate analysis. Let the likelihood be denoted by $x_i | \mu \sim \Gamma(\alpha, \beta)$, so that $\alpha = v/2$, $\beta = v/2\mu$ and let μ^{-1} , be denoted by θ . Then the posterior distribution for μ^{-1} , $(\mu^{-1} | I_T)$ is given by $(\mu^{-1} | I_T) = (\theta | I_T) \propto \text{likelihood} \times \text{prior}$

$$\begin{aligned}
 &= \frac{\beta^\alpha}{\Gamma(\alpha)^T} \times \prod_{i=1}^T (\alpha - 1) e^{-\beta \sum_{i=1}^T x_i} \times \frac{l_0^{k_0}}{\Gamma(k_0)} \theta^{k_0 - 1} e^{-l_0 \theta} \\
 &= \theta^{rv/2} e^{-\frac{\theta v}{2} \sum_{i=1}^T x_i} \theta^{k_0 - 1} e^{-l_0 \theta}, \text{ since } \beta = \frac{v}{2} \theta \text{ and } \alpha = \frac{v}{2} \\
 &\propto \theta^{k_0 + rv/2 - 1} e^{-\theta(l_0 + \frac{v}{2} \sum_{i=1}^T x_i)}.
 \end{aligned}$$

As expected, the posterior distribution of $\mu^{-1} | I_T$, is given by a Gamma distribution, let $\mu^{-1} | I_T \sim \Gamma(k_T, l_T)$. From the above,
 $k_T = k_0 + \frac{rv}{2} - 1 = \frac{1}{2}vn_0 + \frac{rv}{2} + 1$, and

$$l_T = l_0 + \frac{v}{2} \sum_{i=1}^T x_i = \frac{1}{2}vn_0\mu_0 + \frac{v}{2} \sum_{i=1}^T x_i.$$

Using result (10.1) of subsection 10.4.1, the posterior mean m_r for

$$\mu \text{ is given by } m_r = \frac{l_r}{k_r - 1}$$

$$= \frac{\frac{1}{2} n_0 u_0 + \frac{1}{2} \sum_{i=1}^r x_i}{\frac{1}{2} n_0 + r/2 + 1 - 1}$$

$$= \frac{n_0 u_0 + \sum_{i=1}^r x_i}{n_0 + r}$$

i.e. $n_r m_r = n_0 u_0 + \sum_{i=1}^r x_i$, where $n_r = n_0 + r$. Hence the posterior mean is of the form encountered for the general result.

Provided therefore the loss functions are as defined in Section 9.4.2, the Bayesian Decision Scheme $BD(a,b)$ is equivalent to the Cusum Decision Scheme (a,b) and V-mask Graph $\left(n_0, \tan^{-1} \frac{b}{n_0 f} \right)$.

10.4.2 Monitoring of the Normal Variance

An important case of practical interest of the monitoring of the gamma model is that concerning the control of a normal variance μ .

Consider a process described by a normal model, with samples of $v + 1$ observations $y_{i,j}$ each being drawn. Then $y_{i,j} | \mu, \Pi_i \sim \text{IN}(\Pi_i; \mu)$ for $j = 1, \dots, v + 1$. For each sample let X_i define the sample variance, $X_i = \sum_{j=1}^{v+1} (y_{i,j} - \bar{y}_i)^2 / v$, so that X_i will have a gamma distribution. With the prior for μ defined as an Inverse Gamma as in Section 10.4.1, the construction of the monitor follows easily.

The monitor, $\sum_{i=1}^r (x_i - a)$, based on r samples is certainly more accurate than the traditional one based on range (used as estimate of variance) as used in traditional quality control mechanisms.

10.5 Monitoring a Normal Mean with Unknown Variance

Consider the Normal model in which $X_i | \mu \sim \text{IN}(\mu; 1/\theta)$ are independent identically distributed random variables with θ unknown and the parameter μ being monitored. Then from the usual Bayesian conjugate analysis, we define the conditional prior distribution for μ by

$$\mu | \phi, I_0 \sim N(\mu_0; 1/(n_0 \phi)) ,$$

and the marginal distribution of θ by

$$v_0 s_0 \theta \mid I_0 \sim \chi^2_{v_0},$$

where v_0 , s_0 are known constants, with v_0 being the number of degrees of freedom of the χ^2 -distribution.

Then, unconditionally on θ , the prior marginal distribution for μ is given by the t -distribution,

$$\text{i.e. } \mu \mid I_0 \sim t_{v_0} \{ \mu_0 ; s_0 \},$$

where v_0 = number of degrees of freedom of the t -distribution,
 μ_0 = location parameter
 s_0 = precision parameter.

Then the joint distribution of μ and $v_0 s_0 \theta$ will be normal chi-square; standard conjugate analysis will give the following results.

(1) The conditional posterior distribution of μ :

$$\mu \mid \theta, I_r \sim N(m_r, 1/(n_r \theta)),$$

$$\text{where } m_r = \frac{n_0 \mu_0 + \sum_{i=1}^r x_i}{n_0 + r},$$

$$n_r = n_0 + r$$

$$(ii) \quad v_r s_r \phi / I_r \sim \chi^2_{v_r}$$

where

$$v_r = v_0 + r$$

(iii) unconditionally on ϕ , the posterior marginal distribution for μ is given by the t-distribution,

$$i.e. \quad \mu | I_r \sim t_{v_r} \{m_r; \tau_r\},$$

where m_r = location parameter as obtained in (i) above

$v_r = v_0 + r$, as in (ii) above, and

$$\tau_r = s_r / n_r.$$

Thus with m_r and n_r as defined above, the requirements for the general result in Chapter 9 are satisfied. It follows that, with the loss functions as defined previously in Section 9.4.2, the monitoring will be equivalent to a Cusum Decision Scheme, CD(a,b). However, the t-distribution not being a member of the exponential family, there is no equivalence with the SSPRTS.

10.6 Applications : The Case of Monitoring a Normal Variance

A set of data for monthly sales of a commodity referred to as 'Weed' for a period of two years (1955, 1956) is modelled according to the well-known linear growth model as defined in West and Harrison (1989a), Chapter 7.

The variance is being monitored. Using the monitor presented in Chapter 8, the variance is found to be too high at time = month 4, 1956. Then there is an intervention whereby the variance is given a higher estimate, in fact from 8.1 to 15. The resulting graph is shown in Figure 10.1.

Figure 10.2 shows the 1-step ahead forecast error; the graph clearly confirms the increase in variance.

Finally, Figure 10.3 shows the standard error with $2/3$ probability limits. The narrowing down of the limits after intervention at month 4, 1956 reflects clearly the efficiency of the monitor.

FIGURE 10.1. THE VALUE OF INVESTMENT AND DESIGN PROFILES (C)



MEED : 1-STEP AHEAD ERROR

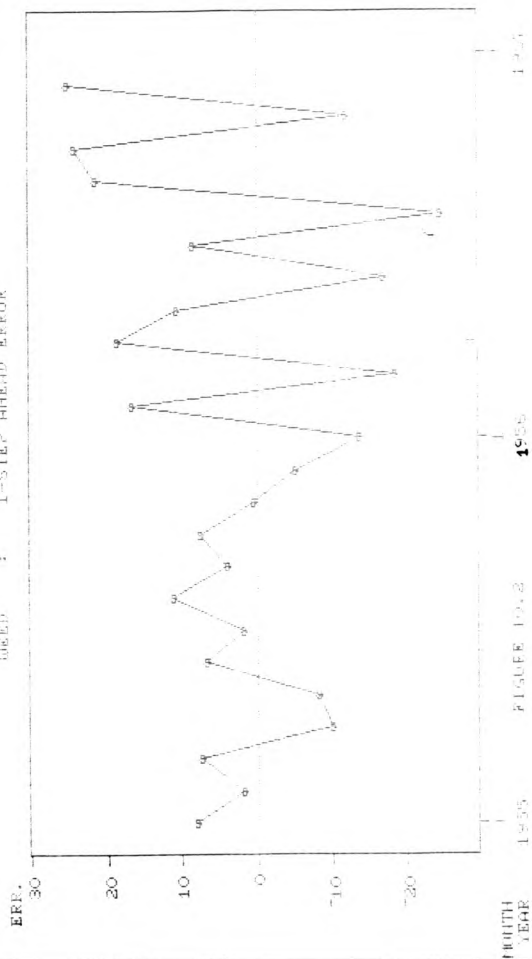


FIGURE 19.2

11. CONCLUSION

In this thesis, new ideas and new results of practical significance dealing with the incorporation and deletion of information on the one hand and with model monitoring and model maintenance on the other hand have been established.

Firstly, basic results for incorporating and deleting information have been derived in a general setting, when the observational variance is known only up to a scalar factor. In particular, expressions for the leverage of a set of observations on the state vector and for moments of the distribution of the state vector with a set of observations deleted have been obtained.

Then, key results regarding the conditional property for normal dynamic linear models have been established. These, in turn, provide an elegant theoretical framework within which a new look at well-known ideas and techniques as well as the development of new results can be carried out. Thus, routine updating and learning procedures, and retrospective analysis are seen as straightforward examples of incorporation of information.

Furthermore, results of practical significance regarding the deletion of information have been developed. Recurrence relationships for the elimination of a single observation are firstly established, which, interestingly, dual those derived for the incorporation of information. These then lead to a simple operation for finite truncated models where the forecasts are based on at most the last k time periods. These results have further been

generalised to the case of the deletion of any set of past observations, with a straightforward procedure for revising distributions. The case of a stochastic variance model has been equally considered.

Particular attention has been given to the subset of discount weighted regression dynamic models since they exhibit very neat procedures and provide a link with static models and least squares procedures.

Bearing in mind that, usually in forecasting systems, the major concern in incorporating and deleting information relates to the present and the recent past, the procedures developed in the thesis offer considerable advantages over those which involve reanalysis of the entire time series history.

The thesis then proceeds with the investigation of model monitoring. Cusum schemes have been critically examined and the two major ways in which cusums operate have been defined. Thereafter, the equivalence of cusums with sequences of SPRTs, particularly for Exponential family models has been examined with elaboration for useful practical cases. However, cusums do not offer a general efficient approach to monitoring.

A link between Cusums and Bayesian Decision Theory has been established for a useful class of linear loss functions. After investigating the Normal mean case, a general result has been established which is directly applicable to many of the important cases such as the Binomial, Poisson and Gamma distributions, the

latter including Normal variance monitoring. It has further been found that these results correspond to the defining characteristics of a Cusum Decision Scheme and the V-mask cusums. In particular, in the latter case, we have an interesting and intuitive result that the distance of the V vertex from the latest plotted point is, in fact, the prior precision in terms of a number of equivalent observations.

The Bayesian Decision approach as well as the SSPRTs are very flexible and easily generalise; they are simple to apply in computer systems. They are to be favoured for general computer model monitoring when compared to cusums.

In the light of the ideas developed and of the results established in the thesis, there are some areas for further research.

1. The methods for incorporating and deleting information can be extended to very general use, applying to non-linear non-normal models in which least squares or linear Bayes procedures are used. They can be equally extended to dynamic generalised linear models, as developed in West and Harrison, 1989, Chapter 14.
2. Regarding specifically the incorporation of information, some interesting cases of practical significance like the case of the incorporation of subjective information and the case of combination of forecasts can further provide useful illustrations of the ideas and results.

3. Concerning the deletion of information, the resulting jackknifed posterior state and various predictive distributions provide the basis for deriving diagnostics such as those advocated by Pettit and Smith (1985), Bernardo (1985) and Johnson and Geisser (1983) and as used in West and Harrison (1991).
4. From a practical point of view, the monitoring of subjective intervention is very important since it may well be ill-founded. The development of monitoring diagnostics would be an area of interest, the more so since they do not necessitate the specification of precise alternative hypotheses. Particular diagnostics relating to functions of the parameters can shed light on the required intervention.

REFERENCES

- AKRAM, M. & HARRISON, P.J. (1983). Applications of Generalised Exponentially Weighted Regression. Research Report 36, Department of Statistics, University of Warwick.
- AMEEN, J.R.M. & HARRISON, P.J. (1984). Discount Weighted Estimation. J.Forecasting, 3, 285-296.
- AMEEN, J.R.M. & HARRISON, P.J. (1985). Normal Discount Bayesian Models. Bayesian Statistics 2, eds. J.M. Bernardo et al.
- ABRAHAM, B. & LEDOLTER, J. (1983). Statistical Models for Forecasting. Wiley.
- BARNARD, G.A. (1959). Control charts and stochastic processes (with discussion). J.R.S.S., B,21, 239-271.
- BARBOSA, E. & HARRISON, P.J. (1989). Variance estimation for multivariate DLMs. Research Report 160, Department of Statistics, University of Warwick.
- BAYES, T. (1763). An essay towards solving a problem in the doctrine of chances. In Biometrika, 45, 1958, 295-315.
- BERGER, J.O. (1985). Statistical Decision Theory and Bayesian Analysis. (2nd edn.). New York, Springer-Verlag.
- BERNARDO, J.M. (1985). Discussion of paper by A.F.M. Smith and L.I. Pettit. In Bayesian Statistics 2. Ed. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, 492-3. Amsterdam: North-Holland; and Valencia University Press.
- BISSELL, A.F. (1969). Cusum techniques for quality control. Applied Statistics 18, 1-30.
- BOX, G.E.P. & JENKINS, G.M. (1976). Time Series Analysis: Forecasting and Control. (2nd edn.). Holden-Day, San Francisco.
- BOX, G.E.P. & TIAO G.C., (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley, Massachusetts.
- BROOK, D. & EVANS, D.A (1972). An approach to the probability distribution of cusum run length. Biometrika, 59, 3, 539-549.
- BROWN, R.G. (1962). Smoothing, Forecasting and Prediction of Discrete Time Series. Prentice-Hall.
- BRUCE, A.G. & MARTIN, R.D. (1989). Leave k-out diagnostics for time series (with discussion). J.R.S.S., B,51,363-424.
- BS24, (1985) BSI Handbook 24. Quality Control. London. British Standards Institute.
- CROSIER, R.B. (1986). A new two-sided cumulative sum quality control scheme. Technometrics, Vol 28, No.3.

- CHATFIELD, C. & COLLINS, A.J. (1980) Introduction to multivariate analysis. Chapman and Hall.
- COOK, R.D. (1977). Detection of Influential Observation in Linear Regression. Technometrics. Vol 19, No.1, Feb. 1977.
- COOK, R.D. & WEISBERG, S. (1982). Residuals and Influence in Regression. Chapman and Hall.
- COX, D.R. & MILLER, H.D. (1970). The Theory of Stochastic Processes. Methuen, London.
- DAVID, A.P. (1979). Conditional Independence in Statistical Theory. J.R.S.S., B, 41, 1-31.
- DE FINETTI, B. (1974, 1975). Theory of Probability (Vols.1 and 2). Wiley, Chichester.
- DE GROOT, M.H. (1971). Optimal Statistical Decisions. McGraw-Hill, New York.
- DRAPER, N.R. & JOHN, J.A. (1981). Influential Observations and Outliers in Regression. Technometrics. Vol.23, No.1, Feb.1981.
- EWAN, W.D. (1963). When and how to use cusum charts. Technometrics. Vol 5, No.1, 1-22.
- EWAN, W.D. & KEMP, K.W. (1960). Sampling inspection of continuous processes with no autocorrelation between successive results. Biometrika. 47, 363-380.
- GILCHRIST, W.G. (1967). Methods of Estimation Involving Discounting. J.R.S.S., B, 29, 355-369.
- GOEL, A.L. & WU, S.M. (1971). Determination of A.R.L. and a contour nomogram for cusum charts to control normal mean. Technometrics. 13, 221-230.
- GRAYBILL, F.A. (1969). Introduction to metrics with applications in statistics. Wadsworth, California.
- HARRISON, P.J. (1965). Short-term Sales Forecasting. J.R.S.S., C, 15, 102-139.
- HARRISON, P.J. (1967). Exponential Smoothing and Short-term Sales Forecasting. Management Science, vol. 13, no. 11, 1967, 821-842.
- HARRISON, P.J. & AKRAM, M. (1982). Generalised Exponentially Weighted Regression and Parsimonious Dynamic Linear Modelling. Research Report 24. Department of Statistics, University of Warwick.
- HARRISON, P.J. & DAVIES, O.L. (1964). The use of cumulative sum (CUSUM) techniques for the control of routine forecasts of produce demand. Oper. Res. 12, 325-333.

- HARRISON, P.J. & JOHNSTON, F.R. Discount Weighted Regression. Research Report 35. Department of Statistics, University of Warwick.
- HARRISON, P.J. & STEVENS, C.F. (1971). A Bayesian Approach to Short-term Forecasting. Oper. Res. Quart., 22, 341-362.
- HARRISON, P.J. & STEVENS, C.F. (1976). Bayesian Forecasting (with Discussion). J.R.S.S. B, vol. 38, no. 3, 205-247.
- HARRISON, P.J. & VEERAPEN, P.P. (1991). A Bayesian decision approach to cusums. Research Report 220. Department of Statistics, University of Warwick.
- HARRISON, P.J. & VEERAPEN, P.P. (1991). Incorporating and deleting information in dynamic models. Research Report 221. Department of Statistics, University of Warwick.
- HARRISON, P.J. & WEST, M. (1991). Dynamic Linear Model diagnostics. Biometrika, 78, ***-***.
- HARRISON, P.J. & WEST, M., & POLE, A. (1987). FAB, a training package for Bayesian forecasting. Warwick Research Report 122, Department of Statistics, University of Warwick.
- HARVEY, A.C. (1981). Time Series Models. Philip Allan, Oxford.
- JEFFREYS, H. (1961). Theory of Probability (3rd edn.). Oxford University Press, London.
- JOHNSON, W. & GEISSER, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. J.Amer.Statist.Assoc. 78, 381, 137-144.
- KEMP, K.W. (1961). The average run length of the cumulative sum chart when a V-mask is used. J.R.S.S. B, 23, 149-153.
- KEMP, K.W. (1962). The use of cumulative sums for sampling inspection schemes. Appl. Statist. 11, 16-31.
- KEMP, K.W. (1971). Formal expressions which can be applied to Cusum charts. J.R.S.S. B, 33, 331-360.
- KOHN, R. & ANSLEY, C.F. (1989). A fast algorithm for signal extraction, influence and cross-validation in state space models. Biometrika, 76, 65-79.
- KULLBACK, S. & LEIBLER, R.A. (1951). On information and sufficiency. Ann. Math. Statist. 22, 1978.
- LINDLEY, D.V. (1956). On the measure of Information provided by an Experiment. Ann.Math.Statist. 27, 986-1005.
- LINDLEY, D.V. (1965). Introduction to Probability and Statistics from a Bayesian viewpoint. (Parts 1 and 2). Cambridge University Press, Cambridge.

- LINDLEY, D.V. (1985). Making Decisions. Second Edition. Wiley, New York.
- LINDLEY, D.V. & SMITH, A.F.M. (1972). Bayes' estimates for the linear model. J.R.S.S., B, 34, 1-41.
- LUCAS, J.M. & CROSIER, R.B. (1982a). Fast initial response for CUSUM quality control schemes: give your CUSUM a head start. Technometrics, vol 24, 199-205.
- LUCAS, J.M. & CROSIER, R.B. (1982b). Robust CUSUM: A robustness study for CUSUM quality control schemes. Communications in Statistics - Theory and Methods. 11, 2669-2687.
- MCCULLAGH, P. & NELDER, J.A. (1989). Generalised Linear Models (2nd edition). Chapman and Hall. London.
- MARDIA, K.V., KENT, J.T., & BIBBY, J.M. (1979). Multivariate Analysis. Academic Press, London.
- PAGE, E.S. (1954). Continuous inspection schemes. Biometrika. 41, 100-115.
- PAGE, E.S. (1961). Cumulative sum charts. Technometrics, vol 3, No. 1, 1-9.
- PETTITT, L.I. & SMITH, A.F.M. (1985). Outliers and Influential Observations in Linear Models. Bayesian Statistics 2, eds. J.M. Bernardo et al.
- QUEEN, C.M. & SMITH, J.Q. (1990). Multiregression dynamic models. Research Report 183, Department of Statistics, University of Warwick.
- QUEEN, C.M. & SMITH, J.Q. (1991). Dynamic graphical models. Research Report 209, Department of Statistics, University of Warwick.
- QUINTANA, J.M. (1987). Multivariate Bayesian forecasting models. Unpublished Ph.D. thesis, Department of Statistics, University of Warwick.
- QUINTANA, J.M. & WEST, M. (1987). Multivariate time series analysis: new techniques applied to international exchange rate data. The Statistician 36, 275-281.
- QUINTANA, J.M. & WEST, M. (1988). Time series analysis of compositional data. In Bayesian Statistics 3. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (Eds.). Oxford University Press.
- SMITH, J.Q. (1988). Decision Analysis. A Bayesian Approach, Chapman and Hall, London.
- SMITH, J.Q. (1979). A Generalisation of the Bayesian steady forecasting model. J.R.S.S., B. Vol.41, 378-387.
- SMITH, J.Q. (1989). Influence diagrams for statistical modelling. The Annals of Statistics, 1989, vol 17, no.2, 654-672.

- SMITH, J.Q. (1990). Statistical principles on graphs (with discussion). In Influence diagrams, belief nets and decision analysis, Ed. R.M. Oliver and J.Q. Smith, 89-120. New York, Wiley.
- WALD, A. (1947). Sequential Analysis. New York, Wiley.
- WEST, M. (1984). Outlier models and prior distributions in Bayesian linear models. J.R.S.S., B, 46, 431-439.
- WEST, M. (1986). Bayesian model monitoring. J.R.S.S., B, 48, 70-78.
- WEST, M. & HARRISON, P.J. (1986). Monitoring and adaptation in Bayesian forecasting models. J.Amer. Statist. Ass. 81, 741-750.
- WEST, M. & HARRISON, P.J. (1989a). Bayesian Forecasting and Dynamic Models. Springer-Verlag, New York.
- WEST, M. & HARRISON, P.J. (1989b). Subjective intervention in formal models. J.Forecasting 8, 33-53.
- WEST, M., HARRISON, P.J. & POLE, A. (1987). BATS: Bayesian analysis of time series. Professional Statistician 6, 43-46.
- WETHERILL, G.B. & BROWN, D.B. (1991). Statistical Process Control. London, Chapman and Hall.
- WOODALL, W.H. (1984). On the Markov Chain approach to the two-sided Cusum procedure. Technometrics. Vol 26, No.1, 41-46.
- WOODWARD, R.H. & GOLDSMITH, P.L. (1964). Cumulative Sum techniques. I.C.I. Monograph No.3 Edinburgh, Oliver and Boyd.
- YOUNG, P.C. (1984). Recursive estimation and time series analysis. Springer-Verlag, Berlin.